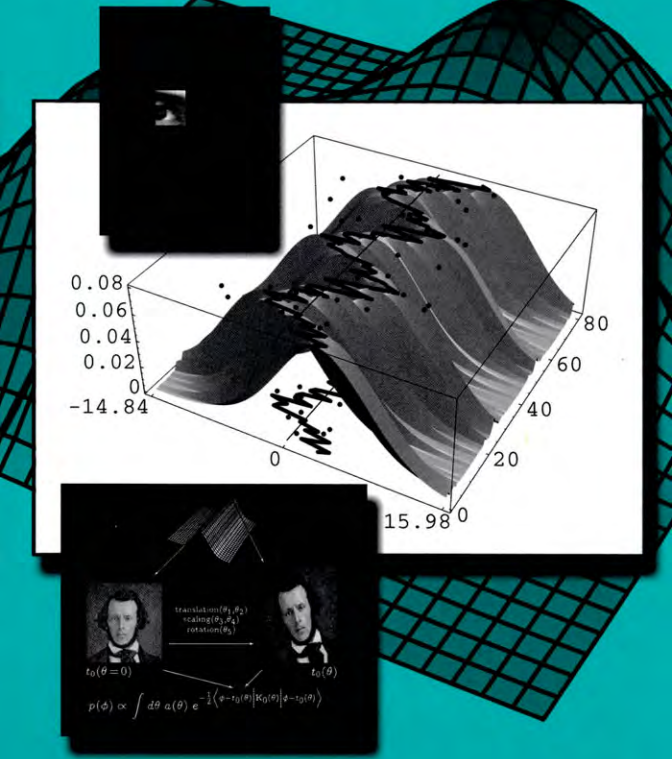
Bayesian Field Theory



Jörg C. Lemm

VISIT…

Bayesian Field Theory

This page intentionally left blank

Bayesian Field Theory

Jörg C. Lemm

The Johns Hopkins University Press
Baltimore and London

© 2003 The Johns Hopkins University Press
All rights reserved. Published 2003
Printed in the United States of America
9 8 7 6 5 4 3 2 1

The Johns Hopkins University Press
2715 North Charles Street
Baltimore, Maryland 21218-4363
www.press.jhu.edu

Figures 7.16, 7.18, 7.20–7.22, and, with minor modifications, parts of Section 7.2 are reprinted from J. C. Lemm, "Inverse time-dependent quantum mechanics," *Physics Letters A*, 276(1–4):19–24, 2000, with permission from Elsevier Science.

Figures 7.1–7.4, 7.6, and, with minor modifications, Sections 7.1.1–7.1.4 and parts of Sections 7.3.1 and 7.3.2 are reprinted from J. C. Lemm and J. Uhlig, "Bayesian inverse quantum theory," *Few-Body Systems*, 29:25–52, 2000, with permission from Springer Verlag.

Figures 7.7–7.15, and, with minor modifications, Sections 5.6, 5.7, and parts of Sections 3.6.2, 3.6.3, 5.5, and 7.1.5 are reprinted from J. C. Lemm and A. Weiguny, "Bayesian reconstruction of approximately periodic potentials for quantum systems at finite temperatures," *The European Journal B*, 20(3):349–366, 2001, with permission from Springer Verlag.

Library of Congress Cataloging-in-Publication Data

Lemm, Jörg C.
Bayesian field theory / Jörg C. Lemm.
p. cm.
Includes bibliographical references and index.
ISBN 0-8018-7220-0 (alk. paper)
1. Bayesian field theory. I. Title.

QC174.85.B38 L45 2002
530.15'95—dc21 2002073958

A catalog record for this book is available from the British Library.

Contents

List of Figures xi

List of Tables xv

List of Numerical Case Studies xvii

Acknowledgments xix

1 Introduction 3

2 Bayesian framework 9
2.1 Bayesian models 9
2.1.1 Independent, dependent, and hidden variables 9
2.1.2 Energies, free energies, and errors 20
2.1.3 Bayes' theorem: Posterior, prior, and likelihood 27
2.1.4 Predictive density and learning 30
2.1.5 Mutual information and learning 33
2.1.6 Maximum A Posteriori Approximation (MAP) 36
2.1.7 Normalization, non–negativity, and specific priors 41
2.1.8 Numerical case study: A fair coin? 46
2.2 Bayesian decision theory 59
2.2.1 Loss and risk 59
2.2.2 Loss functions for approximation 59
2.2.3 General loss functions and unsupervised learning 62
2.2.4 Log–loss and Maximum A Posteriori Approximation . . 63
2.2.5 Empirical risk minimization 64
2.2.6 Interpretations of Occam's razor 67
2.2.7 Approaches to empirical learning 70
2.3 *A priori* information 72
2.3.1 Controlled, measured, and structural priors 72
2.3.2 Noise induced priors 75

3 Gaussian prior factors **85**
3.1 Gaussian prior factor for log–likelihoods 85
3.1.1 Lagrange multipliers: Error functional $E(L)$ 85
3.1.2 Normalization by parameterization: Error functional $E(g)$ 89
3.1.3 The Hessians $\mathbf{H}_L$, $\mathbf{H}_g$ 91
3.2 Gaussian prior factor for likelihoods 92
3.2.1 Lagrange multipliers: Error functional $E(P)$ 92
3.2.2 Normalization by parameterization: Error functional $E(z)$ 94
3.2.3 The Hessians $\mathbf{H}_P$, $\mathbf{H}_z$ 95
3.3 Quadratic density estimation and empirical risk minimization . 96
3.4 Numerical case study: Density estimation with Gaussian specific priors 99
3.5 Gaussian prior factors for general fields 112
3.5.1 The general case 112
3.5.2 Square root of P 114
3.5.3 Distribution functions 115
3.6 Covariances and invariances 116
3.6.1 Approximate invariance 116
3.6.2 Infinitesimal translations 118
3.6.3 Approximate periodicity 120
3.6.4 Approximate fractals 122
3.7 Non–zero means 123
3.8 Regression 126
3.8.1 Gaussian regression 126
3.8.2 Exact predictive density 138
3.8.3 Gaussian mixture regression (cluster regression) 141
3.8.4 Support vector machines and regression 142
3.8.5 Numerical case study: Approximately invariant regression (AIR) 143
3.9 Classification 161

4 Parameterizing likelihoods: Variational methods **167**
4.1 General likelihood parameterizations 167
4.2 Gaussian priors for likelihood parameters 170
4.3 Linear trial spaces 171
4.4 Linear regression 174
4.5 Mixture models 179
4.6 Additive models 179
4.7 Product ansatz 181
4.8 Decision trees 182
4.9 Projection pursuit 183
4.10 Neural networks 183

5 Parameterizing priors: Hyperparameters **187**
5.1 Quenched and annealed prior normalization 187
5.2 Saddle point approximations and hyperparameters 191
5.2.1 Joint MAP 191
5.2.2 Stepwise MAP 192
5.2.3 Pointwise approximation 192
5.2.4 Marginal posterior and empirical Bayes 193
5.2.5 Some variants of stationarity equations 195
5.3 Adapting prior means 197
5.3.1 General considerations 197
5.3.2 Density estimation and nonparametric boosting 198
5.3.3 Unrestricted variation 199
5.3.4 Regression 199
5.4 Adapting prior covariances 202
5.4.1 General case 202
5.4.2 Automatic relevance detection 202
5.4.3 Local masses and gauge theories 203
5.4.4 Invariant determinants 204
5.4.5 Regularization parameters 207
5.4.6 Cross–validation 210
5.5 Integer hyperparameters 215
5.6 Hyperfields 216
5.7 Auxiliary fields 221
5.8 Non–quadratic potentials 224

6 Mixtures of Gaussian prior factors **229**
6.1 Multimodal energy surfaces 229
6.2 Prior mixtures for density estimation 231
6.3 Numerical case study: Prior mixtures for density estimation . . 231
6.4 Prior mixtures for regression 239
6.4.1 General framework 239
6.4.2 High and low temperature limits 240
6.4.3 Equal covariances and the ferromagnet 242
6.4.4 Some analytical calculations for mixture models 244
6.5 Local mixtures 248
6.6 Numerical case study: Image completion 249

7 Bayesian inverse quantum theory (BIQT) **257**
7.1 Bayesian inverse quantum statistics (BIQS) 258
7.1.1 The likelihood model of quantum theory 258
7.1.2 Prior models for potentials 263
7.1.3 Numerical case study: Inverse quantum statistics 269
7.1.4 Classical approximation 273

7.1.5 Numerical case study: Reconstruction of perturbed periodic potentials using hyperfields and auxiliary fields . 276
7.2 Bayesian inverse time–dependent quantum theory (BITDQ) . . 288
7.2.1 Quantum time series 288
7.2.2 Numerical case study: Time–dependent systems 291
7.3 Bayesian inverse many–body theory 299
7.3.1 Systems of fermions 299
7.3.2 Bayesian inverse Hartree–Fock theory (BIHF) 301
7.3.3 Numerical case study: Inverse Hartree–Fock theory . . . 304

8 Summary **309**

A *A priori* information and *a posteriori* control **313**

B Probability, free energy, energy, information, entropy, and temperature **323**
B.1 Statistics and statistical mechanics 323
B.2 Probability 325
B.2.1 Normalization factors or partition sums 325
B.2.2 Log–probabilities, bit numbers, and free energies 326
B.3 Random variables 326
B.3.1 Averages 327
B.3.2 Information 327
B.3.3 Energy 328
B.4 Temperature and external fields 330
B.4.1 Maximum entropy and Boltzmann–Gibbs distributions . 331
B.4.2 Annealing methods 332
B.4.3 Generating functions 332
B.5 Conditional probabilities and disordered systems 339

C Iteration procedures: Learning **345**
C.1 Numerical solution of stationarity equations 345
C.2 Learning matrices 348
C.2.1 Learning algorithms for density estimation 348
C.2.2 Linearization and Newton algorithm 349
C.2.3 Massive relaxation 350
C.2.4 Gaussian relaxation 354
C.2.5 Inverting in subspaces 354
C.2.6 Boundary conditions 355
C.3 Initial configurations and kernel methods 357
C.3.1 Truncated equations 357
C.3.2 Kernels for L 359
C.3.3 Kernels for P 361

Bibliography **365**

Index **403**

This page intentionally left blank

Figures

2.1 The basic model 11
2.2 Graphical representation of the probabilistic model 13
2.3 Schematic representation of nonparametric and parametric models 18
2.4 Additional noise processes 19
2.5 Graphical representation of Bayes' theorem 30
2.6 Graphical representation of learning 31
2.7 Evolution of the predictive density 32
2.8 Graphical representation of a factorial prior 37
2.9 The decay of mutual information during learning 38
2.10 The coin example 1 (uniform prior, $n = 1$) 51
2.11 The coin example 2 (uniform prior, $n = 2$) 52
2.12 The coin example 3 (uniform prior, $n = 3$) 53
2.13 The coin example 4 (uniform prior, $n = 10$) 54
2.14 The coin example 5 ('expert' prior, $n = 10$) 55
2.15 The coin example 6 ('Las Vegas' prior, $n = 10$) 56
2.16 The coin example 7 (posteriors for three different priors, varying k, $n = 10$) 57
2.17 The coin example 8 (evolution of the three posteriors) 58
2.18 Bayesian approach vs. Frequentist hypothesis testing 68
2.19 Bayesian approach vs. Frequentist hypothesis testing: Example 69
2.20 Example of the influence of stopping rules 70
2.21 Measured and structural dependencies 75
2.22 Output averaging with multiple inputs 79
2.23 Noise and learning 80

3.1 Density estimation 1 (2 data points with Gaussian prior factor for L) 102
3.2 Density estimation 2 (2 data points with Gaussian prior factor for P) 103
3.3 Density estimation 3 (2 data points with $(-\Delta^3)$ Gaussian prior factor for L) 104

3.4 Density estimation 4 (2 data points with $(-\Delta^3)$ Gaussian prior factor for P) . . . 105
3.5 Density estimation 5 (50 data points with Gaussian prior factor for L) . . . 108
3.6 Density estimation 6 (True density for examples) . . . 109
3.7 Empirical and conditional empirical density . . . 109
3.8 Regression function for examples . . . 110
3.9 Comparison of iteration schemes and different initializations . . 111
3.10 Three robust error functions . . . 143
3.11 AIR 1: Scheme of nonparametric (approximately invariant) Bayesian regression . . . 146
3.12 AIR 2: Smooth Gaussian likelihood (negative Laplacian) . . . 147
3.13 AIR 3: Smooth Gaussian likelihood (truncated RBF) . . . 148
3.14 AIR 4: Smoothness (negative Laplacian): Prior covariance and equivalent kernel . . . 149
3.15 AIR 5: Smoothness (truncated RBF): Prior covariance and equivalent kernel . . . 149
3.16 AIR 6: Smooth regression with a negative Laplacian prior . . . 150
3.17 AIR 7: Smooth regression with a truncated RBF prior 1 . . . 151
3.18 AIR 8: Smooth regression with a truncated RBF prior 2 . . . 152
3.19 AIR 9: Approximately periodic Gaussian likelihood with negative Laplacian . . . 153
3.20 AIR 10: Approximately periodic Gaussian likelihood with RBF prior . . . 154
3.21 AIR 11: Approximate periodicity 1: Prior covariance and equivalent kernel . . . 155
3.22 AIR 12: Approximate periodicity 2: Prior covariance and equivalent kernel . . . 155
3.23 AIR 13: Approximately periodic regression 1 . . . 156
3.24 AIR 14: Approximately periodic regression 2 . . . 157
3.25 AIR 15: Approximately fractal Gaussian likelihood . . . 158
3.26 AIR 16: Approximate fractal: Prior covariance and equivalent kernel . . . 159
3.27 AIR 17: Approximately fractal regression . . . 160
3.28 Graphical representation of an indirect learning model . . . 163

5.1 Joint posterior density of ϕ and λ . . . 209
5.2 Non-quadratic potentials . . . 226

6.1 Density estimation 7 (The true density and templates used for the examples) . . . 234
6.2 Density estimation 8 (50 data points with a Gaussian mixture prior) . . . 235

6.3 Density estimation 9 (50 data points with a Gaussian mixture prior and large λ) 236
6.4 Density estimation 10 (50 data points with a Gaussian mixture prior and small λ) 237
6.5 Density estimation 11 (Gaussian mixture prior: Example of an approximately stable solution) 238
6.6 Gaussian mixture prior and the ferromagnet 244
6.7 Gaussian mixture prior and the ferromagnet (3D plot) 245
6.8 Joint posterior density of ϕ and β 246
6.9 Joint posterior density of ϕ and β (different data) 247
6.10 Exact and MAP mixture coefficients 248
6.11 Image completion 1: Training data 251
6.12 Image completion 2: Gaussian mixture prior 253
6.13 Image completion 3: Prior with hyperparameters 254
6.14 Image completion 4: Low temperature solution 255
6.15 Image completion 5: High temperature solution 256

7.1 BIQS 1: Approximation of symmetric potentials 1 270
7.2 BIQS 2: Approximation of symmetric potentials 2 270
7.3 BIQS 3: Approximation of symmetric potentials 3 (overfitting) 271
7.4 BIQS 4: Approximation of symmetric potentials 4 (Gaussian mixture prior) 272
7.5 BIQS 5: Scheme of Bayesian inverse quantum statistics 274
7.6 BIQS 6: Classical approximation of a symmetric potential . . . 275
7.7 BIQS 7: Reconstruction of an approximately periodic potential with smoothness prior 283
7.8 BIQS 8: Reconstruction of an approximately periodic potential with periodic reference potential 1 284
7.9 BIQS 9: Reconstruction of an approximately periodic potential with periodic reference potential 2 284
7.10 BIQS 10: Reconstruction of an approximately periodic potential with periodic reference potential 3 285
7.11 BIQS 11: Reconstruction of an approximately periodic potential with approximately periodic inverse covariance 285
7.12 BIQS 12: Reconstruction of an approximately periodic potential with local switching 1 286
7.13 BIQS 13: Generation of new trial configurations 286
7.14 BIQS 14: Reconstruction of an approximately periodic potential with local switching 2 287
7.15 BIQS 15: Reconstruction of an approximately periodic potential with local hyperfield 287
7.16 BITDQ 1: Time evolution of an unobserved particle 293
7.17 BITDQ 2: Time evolution of an observed particle 1 294
7.18 BITDQ 3: Time evolution of an observed particle 2 295

7.19 BITDQ 4: The quantum Zeno effect 296
7.20 BITDQ 5: Numerical reconstruction of a potential 296
7.21 BITDQ 6: Empirical transition densities 297
7.22 BITDQ 7: Conditional empirical transition densities 297
7.23 BITDQ 8: Scheme of Bayesian inverse time–dependent quantum theory 298
7.24 BIHF 1: Two–particle data (case A) 306
7.25 BIHF 2: Single–particle data (case B) 307
7.26 BIHF 3: Scheme of Bayesian inverse Hartree–Fock approximation 308

C.1 Initial guesses for P 363
C.2 Nonparametric classification: An example 364

Tables

1.1 Overview 8

2.1 The basic quantities 15
2.2 The basic probabilities 16

3.1 Constraints for specific choices of ϕ 113
3.2 Some special cases of density estimation 165

4.1 Summary of stationarity equations 171
4.2 Some possible likelihood parameterizations 186

5.1 Some variants of implicit and explicit implementation of *a priori* information 201

7.1 Examples of density operators for quantum systems 259
7.2 Summary of basic prior energy variants 277

B.1 Transformations of probability and averages of the corresponding canonical random variables 330
B.2 Generating functions for zero field 339
B.3 Generating functions for nonzero fields 340

C.1 Learning matrices 358

This page intentionally left blank

Numerical Case Studies

1 A fair coin? (2.1.8) 46

2 Density estimation with Gaussian specific priors (3.4) 99

3 AIR: Approximately invariant regression (3.8.5) 143

4 Prior mixtures for density estimation (6.3) 231

5 Image completion (6.6) 249

6 BIQS: Bayesian inverse quantum statistics (7.1.3) 269

7 BIQS using hyperfields and auxiliary fields: Reconstruction of perturbed periodic potentials (7.1.5) 276

8 BITDQ: Bayesian inverse time–dependent quantum theory (7.2.2) 291

9 BIHF: Bayesian inverse Hartree–Fock theory for many–body quantum systems (7.3.3) 304

This page intentionally left blank

Acknowledgments

It is a great pleasure to thank Achim Weiguny for his support throughout this work, his careful reading of the manuscript, and him and Joerg Uhlig for the enjoyable collaboration. Essential parts of this work are based on the following articles of the author and these two collaborators: [291–300, 303–306].

I am also very grateful to Tomaso Poggio for the opportunity to work in the most stimulating atmosphere of his group and to him and the members of his group, in particular Federico Girosi, for many fruitful discussions. Many of the ideas worked out in this paper originate from that visit at the Massachusetts Institute of Technology which was made possible by a Postdoctoral Fellowship (Le 1014/1-1) from the Deutsche Forschungsgemeinschaft and a NSF/CISE Postdoctoral Fellowship.

Special thanks go to Gernot Münster, Manfred Stingl, Christian Wieczerkowski, Klaus Pinn and all the other members of the Institut für Theoretische Physik for the many interesting discussions about field theory in seminars and at lunch time.

And finally I want to express my sincere thanks to my wife Anja, as well as to my son Marius and my daughter Julia, for their support and patience during the time this work had to be completed.

Abstract

Bayesian field theory stands for a nonparametric Bayesian approach to learning from observational data. Based on the principles of Bayesian statistics, a particular Bayesian field theory is defined by the combination of two models: 1) A likelihood model, providing a probabilistic description of the measurement process for the observational ('training') data, characterizing the area for which application of the theory is intended and 2) a prior model, providing the information which is necessary to generalize from training to non–training data. The particular likelihood models discussed in this book are those of general density estimation, Gaussian regression, clustering, classification, or pattern recognition, as well as specific models of inverse quantum theory. These models represent the most common problem types of empirical learning, studied in many disciplines like applied statistics, artificial intelligence, and physics. Prior models have to implement problem–typical hard constraints, like normalization and non–negativity for probabilities, and also all the vague and probabilistic *a priori* knowledge available for a specific task. Due to their flexibility, nonparametric approaches rely much more on an adequate implementation of *a priori* information than typical parametric methods. The book therefore intends to provide a tool box to deal with *a priori* information in nonparametric models. The nonparametric prior models treated in this book include Gaussian processes, mixtures of Gaussian processes, non–quadratic potentials, as well as so–called hyperparameters, hyperfields and auxiliary fields, all of which can be seen as specific statistical field theories. For applications a collection of practical methods is developed to adapt prior models. In particular, the adaption of mean functions and covariance operators of Gaussian process components is discussed in detail. Bayesian field theories are typically non–Gaussian and have thus to be solved numerically. Thanks to increasing computational resources the class of non–Gaussian Bayesian field theories of practical interest which are numerically feasible is steadily growing. Where a direct numerical solution is computationally too demanding, a variety of approximation methods based on variational techniques is available.

Bayesian Field Theory

This page intentionally left blank

Chapter 1

Introduction

Due to increasing computational resources, the last decade has seen a rapidly growing interest in applied empirical learning problems. They appear, for example, as density estimation, regression or classification problems and include, just to name a few, image reconstruction, speech recognition, time series prediction, object recognition, and also inverse problems in physics.

Many disciplines, like applied statistics, artificial intelligence, computational and statistical learning theory, statistical physics, and also psychology and biology, contributed in developing a variety of learning algorithms. Those include, for example, smoothing splines [400,443,455,507], regularization and kernel approaches [482,483,492], support vector machines [85,493,494], generalized additive models [203], projection pursuit regression [227], expert systems [50,127,178,205,230], decision trees [64], neural networks [52,209,433], and graphical models [241,245,246,282,386]. Recently, much work has been devoted to the comparison and unification of methods arising from different disciplines. (For an overview and comparison of methods see, for example, [124,340,400].) In particular, it turned out that the Bayesian approach to statistics provides a useful unifying framework for empirical learning problems [27,33,38,61,65,73,95,150,234,237,238,280,281,308,314,370,371,424,457,517].

Empirical learning, i.e., the problem of finding underlying general laws from observations, represents a typical inverse problem and is usually ill–posed in the sense of Hadamard [216,221,222,262,315,353,478,482,483,492,494]. It is well known that a successful solution of such problems requires additional *a priori* information. It is *a priori* information which controls the generalization ability of a learning system by providing the link between available empirical training data and unknown outcome in future 'test' situations. In Bayesian approaches predictions of test data are based on the combination of a probabilistic model for the measured training data with a probabilistic model representing the available *a priori* information.

In this book we will study general density estimation problems. Those

include, as special cases, regression, classification, certain types of clustering, and inverse quantum theory. In *density estimation* the functions of interest are the (sampling) probability densities $p(y|x,\phi)$ for outcome y (output, 'dependent part of data') under condition x (input, 'independent part of data', under explicit control) and unknown state of Nature ϕ. Considered as function of ϕ, for fixed y, x, the function $p(y|x,\phi)$ is known as *likelihood* function (following [131]). A Bayesian approach to density estimationis based on a probabilistic model for likelihoods $p(y|x,\phi)$. We will concentrate on situations where y and/or x are real variables, possibly multi–dimensional.

In *nonparametric* approaches the likelihood is determined by a function or 'field' ϕ. A function $\phi(x,y)$ can directly represent the likelihood function $p(y|x,\phi)$ itself. In that case ϕ stands for the collection of numbers $0 \leq p(y|x,\phi)$, for all x and all y, with $\int dy\, p(y|x,\phi) = 1$, or $\sum_y p(y|x,\phi) = 1$, respectively, for all x. Similarly, ϕ can represent the log–likelihoods $\ln p(y|x,\phi)$ or some other function of the likelihoods, possibly nonlocal with respect to x and/or y. In nonparametric approaches the dimension of ϕ is infinite, if the number of values which the variables x and/or y can take is infinite, as it is the case for real x and/or y.

We speak of *parametric* approaches if the likelihood is parameterized by some (e.g., real) parameters, and not by a whole function $\phi(x,y)$. Denoting such parameters of the likelihood by ξ (instead of ϕ) to distinguish parametric from nonparametric approaches, we write $p(y|x,\xi)$ for parameterized likelihoods. For example, a parametric model, parameterized by a vector $\xi = (\xi_0,\xi_1)$, is obtained by considering only likelihoods $p(y|x,\xi)$ of Gaussian form with known variance and a linear regression function $\int dy\, y\, p(y|x,\xi) = \xi_0 + \xi_1 x$. In a more complex parametric model the regression function may be modeled by a neural network. (Note that because even nonparametric models finally have to be discretized for actual numerical calculations and the number of parameters can be large in parametric models, like in neural networks, the distinction between parametric and nonparametric approaches is not sharp in practice.)

While the problem of reconstructing (learning) a general probability density $p(y|x,\phi)$, or $p(y|x,\xi)$, respectively, is known as density estimation, the special case of a problem with discrete y–variable is also called a *classification problem.* Restricting the likelihoods to Gaussian probability densities $p(y|x,\phi)$ with fixed variance leads to *Gaussian regression problems.* Similarly, restricting the likelihoods to some non–Gaussian form, with only the mean unknown for each value of x, leads to *non–Gaussian regression.* For regression problems the aim is to find an optimal regression function $\phi(x)$. Furthermore, adapting a mixture of Gaussians allows *soft clustering* of data points, while extracting relevant features from the predictive density $p(y|x,\text{data})$ is the Bayesian analogue of *unsupervised learning.* A special class of density estimation problems are *inverse problems in quantum physics* where ϕ represents an unknown potential to be determined from observational data [298,304,305,489]. Special

emphasis will be put on the explicit and flexible implementation of *a priori* information using, for example, mixtures of Gaussian prior processes with adaptive, non–zero mean functions for the mixture components.

We will focus mainly on nonparametric approaches (also sometimes called 'form free' or 'free form' approaches) which are formulated directly in terms of the likelihood values of interest. Parametric methods, on the other hand, impose implicit restrictions on the likelihoods which can be extremely difficult to relate to available *a priori* knowledge. Combined with a Bayesian framework, nonparametric approaches typically allow a very flexible and interpretable implementation of *a priori* information in terms of stochastic processes [111, 142, 221, 297, 299, 309, 364, 385, 491, 507, 521, 522].

Nonparametric Bayesian methods can easily be adapted to different learning situations and have therefore been applied to a variety of empirical learning problems, including regression, classification, density estimation, and inverse quantum problems [298,303–305,363,489,522]. Technically, they are related to kernel and regularization methods which often appear in the form of a roughness penalty approach [39, 183, 198, 259, 263, 332, 401, 456, 483, 492, 494, 507]. Computationally, working with stochastic processes, or discretized versions thereof, is more demanding than, for example, fitting a small number of parameters. This holds especially for such applications where one can not take full advantage of the convenient analytical features of Gaussian processes. Nevertheless, it seems to be the right time to study nonparametric Bayesian approaches also for non–Gaussian problems, as they become computationally feasible, at least for low dimensional systems, and they provide a well defined basis for further approximations even in such cases where they are not directly solvable.

From a physicist's point of view functions depending on continuous variables, like $\phi(x, y) = p(y|x, \phi)$, are also called a 'field'. A Bayesian approach requires to implement the *a priori* information which is available for the possible likelihood models. In a nonparametric model such *a priori* informations are represented by a stochastic process $p(\phi)$ which is directly formulated in terms of the field values $\phi(x, y)$, e.g., $p(\phi) = p(\phi(x, y), x \in X, y \in Y)$. (In the parametric case, discussed in Chapter 4, one works instead with probability densities $p(\xi)$ for parameters ξ.) Models defined by stochastic processes, like in our case $p(\phi)$, are typically called 'field theories' in physics. Since from that point of view nonparametric Bayesian models represent specific (statistical) field theories, we will call them *Bayesian field theories.*[1] Based on this relation to field theory the motivation for writing this book was twofold:

1. To make use of the connection between nonparametric Bayesian approaches and statistical field theories to develop and study systematic

[1] We may also remark that statistical field theories, comprise quantum mechanics and quantum field theory in their Euclidean formulation [232, 284, 541]. Thus they can, at least technically, be related to nonparametric Bayesian approaches.

methods for incorporating *a priori* information in nonparametric approaches. The different field theories corresponding to different implementations of *a priori* information are the topic of Chapters 3 – 6.

2. To apply the nonparametric Bayesian framework to a genuine physical problem. As such an application, a new method for solving inverse quantum problems is presented in Chapter 7.

Because Gaussian integrals can be solved analytically, Gaussian processes or, in the language of physics, (generalized) free fields, are very attractive starting points for nonparametric learning. Unfortunately, only the case of Gaussian regression is completely Gaussian. For general density estimation problems the likelihood terms are non–Gaussian, and even for Gaussian prior processes additional non–Gaussian restrictions have to be included to ensure non–negativity and normalization of densities. Hence, in the general case, density estimation is a non–Gaussian theory, or, in the language of physics, an interacting field theory.

As it is well known from physics, a continuum limit for non–Gaussian theories, based on the definition of a renormalization procedure, can be highly nontrivial to construct. (See [6,45] for a renormalization approach to density estimation.) We will in the following not discuss such renormalization procedures but focus more on practical, numerical learning algorithms, obtained by discretizing the problem (typically, but not necessarily in coordinate space). This is similar to what is done in lattice field theories.

Gaussian problems live effectively in a space with dimensions not larger than the number of training data. This is not the case for non–Gaussian problems. Hence numerical implementations of learning algorithms for non–Gaussian problems require to discretize the functions of interest. This can be computationally challenging. For low dimensional problems, however, many non–Gaussian models are nowadays solvable on a standard PC. Examples include predictions of one–dimensional time series or the reconstruction of two–dimensional images. Higher dimensional problems require approximations, like projections into lower dimensional subspaces or other variational approaches. Indeed, it seems that most solvable high dimensional problems live effectively in some low dimensional subspace. There are also approaches to classification where non–negativity and normalization constraints are fulfilled automatically. In that case, the calculations can still be performed in a space of dimension not larger than the number of training data. Contrasting Gaussian models, however, the equations to be solved are then typically nonlinear.

Finally, the book wants to show that Bayesian approaches can, in a straightforward way, be adapted to inverse quantum problems by using likelihood models as defined by the axioms of quantum theory. A typical problem of this kind is the reconstruction of potentials from quantum mechanical measurements which will be discussed in some detail.

Summarizing, the book discusses a unifying framework for nonparametric Bayesian models, called Bayesian field theory, to learn a function of one or more continuous variables and having in mind especially non–Gaussian models. A large variety of Bayesian field theories can be constructed by combining specific likelihood models with specific functional priors (see Table 1.1). The resulting flexibility of nonparametric Bayesian approaches and their possibility to treat *a priori* information explicitly are probably their main advantages.

The book is organized as follows: Chapter 2 summarizes the Bayesian framework as needed for the subsequent chapters. Basic notations are defined, an introduction to Bayesian decision theory is given, and the role of *a priori* information is discussed together with the basics of the Maximum A Posteriori Approximation (MAP) as well as the specific constraints for density estimation problems. Gaussian prior processes, being the most commonly used prior processes in nonparametric statistics, are treated in Chapter 3. In combination with Gaussian prior models, Chapter 3 also introduces the likelihood models of density estimation (Sections 3.1, 3.2, 3.5), Gaussian regression and clustering (Section 3.8), and classification (Section 3.9). Notice, however, that all these likelihood models can also be combined with the more elaborate prior models discussed later in Chapters 5 and 6. Parametric approaches, useful if a numerical solution of a full nonparametric approach is not feasible, are the topic of Chapter 4. Hyperparameters, parameterizing prior processes and making them more flexible, are considered in Chapter 5. Two possibilities to go beyond Gaussian processes, non–quadratic prior potentials and prior mixture models, are presented in Section 5.8 and Chapter 6. The Bayesian approach to inverse quantum problems is discussed in Chapter 7, including inverse quantum statistics (Section 7.1), time–dependent inverse problems (Section 7.2), and many–body systems (Section 7.3).

Using a nonparametric formulation, the role of *a priori* information is more prominent than in standard parametric Bayesian approaches. For example, within a parametric model the influence of *a priori* information becomes typically (but not necessarily) negligible when the number of data becomes very large. This is not true anymore in a nonparametric framework. A general discussion of the role of *a priori* information from a nonparametric point of view has therefore been included in the book as Appendix A.

Appendix B presents an overview on basic concepts of statistical mechanics like free energy, energy, information, entropy, and temperature, which are all related to probability. Appendix C focuses on learning algorithms, i.e., on methods to solve the stationarity equations resulting from the Maximum A Posteriori Approximation for general density estimation problems. Typical kernel methods for density estimation appear as the first iteration step of a general iteration procedure.

Likelihood Model	**Prior Model**
describes	
measurement process (Chap. 2)	generalization behavior (Chap. 2)
is determined by	
parameters (Chaps. 3 and 4)	hyperparameters (Chap. 5)
Examples include	Variants are
density estimation(Sects. 3.1–3.3, 6.2)	hard constraints (Chap. 2)
regression (Sects. 3.8, 6.4) (and classical inverse problems)	Gaussian prior factors (Chap. 3)
classification (Sect. 3.9)	non–quadratic potentials (Sect. 5.8)
inverse quantum theory (Chap. 7)	mixture of Gaussians (Chap. 6)
The role of *a priori* information is discussed in Section 2.3 and Appendix A	
An overview over some notations used in statistical mechanics can be found in Appendix B	
Learning algorithms are treated in Appendix C	

Table 1.1: Overview. A Bayesian approach is based on the combination of two models, a likelihood model, describing the measurement process used to obtain the training data, and a prior model, enabling generalization to non–training data. Parameters of the prior model are commonly called hyperparameters. In nonparametric approaches the collection of all values of the likelihood function are taken as 'parameters' or primary degrees of freedom. A nonparametric Bayesian approach for likelihoods depending on one or more real variables is in this book called a Bayesian field theory.

Chapter 2

Bayesian framework

2.1 Bayesian models

2.1.1 Independent, dependent, and hidden variables

Looking for a scientific explanation of some phenomena means to search for a causal model which relates the relevant observations under study. To define a causal structure, observable (or visible) variables are separated into dependent variables ('measured effects', 'answers') and independent variables ('controlled causes', 'questions'). Dependent variables represent the results of measurements the theory is aiming to explain. Independent variables are assumed to be under external (stochastic) control and characterize the situation and the kind of measurements performed.

Theories typically introduce concepts which are not directly observable. The formulation of causal relations can often be considerably simplified by introducing hidden or latent variables ('hidden causes', 'state of Nature'). For example, the assumption that observational data are sampled from a stationary probability distribution allows one to infer from data the (hidden) parameters of that underlying distribution. The reconstructed distribution can then be used to predict new data. Similarly, symptoms (visible variables) are grouped in medicine into syndromes and diseases (latent variables) which are easier to handle than collections of individual symptoms.

Hence we will consider the following three groups of variables

1. observable (visible) independent variables x ('controlled causes'),

2. observable (visible) dependent variables y ('measured effects'),

3. not directly observable (hidden, latent) variables ϕ ('(possible, pure) state of Nature').

A probabilistic model for these variables is defined by giving a joint probability density $p(x, y, \phi)$ [241, 245, 282, 386, 388, 416] A joint density is best given by decomposing it into smaller factors. Indeed, using $p(A, B) = p(A)p(B|A) = p(B)p(A|B)$ a joint density can be factorized in various ways. Each of these factorizations corresponds to an ordering of the variables (causal structure or 'time'). For instance,

$$\begin{aligned} p(x, y, \phi) &= p(y|x, \phi)\, p(\phi|x)p(x) \\ &= p(\phi|x, y)\, p(y|x)p(x) \\ &= p(y|x, \phi)\, p(x|\phi)p(\phi) \end{aligned} \tag{2.1}$$

are possibilities to factorize the joint density in the above three variables. The given examples correspond to the orderings (x, ϕ, y), (x, y, ϕ), or (ϕ, x, y) respectively. Clearly, in case x, y, ϕ are vector variables, such causal orderings can also mix components of different variables. Note that these factorizations are completely general and do all describe the same probabilistic model. They just reflect different causal interpretations of the model or different ways to specify the model.

A specific probabilistic model is obtained by defining all its conditional and unconditional factors. Of particular practical importance are specifications which implement additional independence assumptions. For example, choosing $p(\phi|x) = p(\phi)$, one models ϕ and x as (stochastically) independent variables. This can drastically reduce the dimensionality of a problem. As such additional simplifications arising from independent variables are in general not visible in all of the possible factorizations into conditional probabilities, for restricted models some specific factorizations are better suited than others. For example, the independence assumption $p(C|A, B) = p(C|A)$, which allows to write $p(A, B, C) = p(A)p(B|A)p(C|A)$, is hidden in the factorization $p(A, B, C) = p(C)p(B|C)p(A|B, C)$.

Unconditional factors, like $p(x)$, are called *a priori* densities. They represent the densities of variables which have been selected as independent variables. Note that the word 'independent' appears in two slightly different meanings. A pair or a set of variables is (stochastically) independent if their joint density factorizes in unconditional factors, $p(A, B) = p(A)p(B)$. In contrast, a single variable is considered independent in a given factorization if it appears within an unconditional factor. For example, when writing $p(A, B) = p(A)p(B|A)$ then A has been chosen as the independent variable. Clearly, each factorization results in at least one independent variable. Two variables can be chosen as independent variables if they are also stochastically independent.

The above characterization of x, y, ϕ as independent, dependent, and hidden variables translates into the following basic model

$$p(x, y, \phi) = p(y|x, \phi)\, p(\phi|x)p(x). \tag{2.2}$$

A graphical representation of (2.2) is shown in Fig. 2.1a. More complex models can be obtained if y and ϕ are vector variables. If x is given, it is not necessary

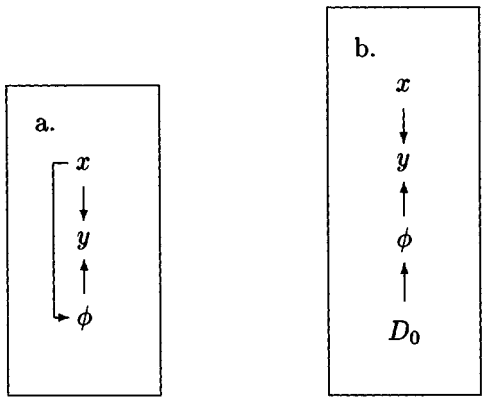


Figure 2.1: The basic model. a. Graphical representation of the probabilistic model (2.2) factorizing according to $p(x, y, \phi) = p(y|x, \phi)\ p(\phi|x)p(x)$ or $p(y, \phi|x) = p(y|x, \phi)\ p(\phi|x)$, for given x. b. Graphical representation of the probabilistic model (2.3) factorizing according to $p(x, y, \phi|D_0) = p(y|x, \phi)\ p(\phi|D_0)p(x)$ or $p(y, \phi|x, D_0) = p(y|x, \phi)\ p(\phi|D_0)$ for given x. The variable D_0 represents the available *a priori* information on the distribution of the hidden variables ϕ.

to include a probabilistic model $p(x)$ for x. Then it is sufficient to work with the conditional model $p(y, \phi|x) = p(y|x, \phi)\, p(\phi|x)$.

The role of x in model (2.2) is twofold. First, setting x controls the preparation of the state ϕ as specified by $p(\phi|x)$. Second, it influences simultaneously with ϕ the measured outcome y according to $p(y|x, \phi)$. Often it is possible to treat these two influences separately, so we can then write $p(\phi|D_0)$ instead of $p(\phi|x)$. For that case, let us define

4. the *a priori* information D_0 on ϕ as that part of the independent visible variables which determines the *a priori* distribution $p(\phi|D_0)$ (*prior*) of the hidden variables ϕ.

In this notation we obtain for model (2.2) conditional on x, D_0, i.e., for the joint prior of y and ϕ (see Fig. 2.1b.)

$$p(y, \phi|x, D_0) = p(y|x, \phi)\, p(\phi|D_0). \tag{2.3}$$

The (data generating or predictive) density for y under condition x in state ϕ, i.e., $p(y|x, \phi)$, considered as function of ϕ, is also known as ***likelihood*** function. So we have in words

$$\text{joint prior} = \text{likelihood} \times \text{prior}. \tag{2.4}$$

In the following, probabilistic models like (2.3) will serve as technical tools for empirical learning.

Note that D_0 does not represent the complete *a priori* information but only that part which determines the prior $p(\phi|D_0)$ for the hidden variables ϕ. (The prior is either a probability, a probability density or, for functions ϕ, a stochastic process, respectively.) The complete *a priori* information also includes the selection of variables, the factorization model (2.2) itself, the specification of the *a priori* probabilities (densities) for all independent variables x, and the conditional probabilities (densities) for all combinations of the dependent variables x and the ϕ as far as they are not yet completely determined by ϕ. Therefore,

5. the part of *a priori* information not contained in D_0 will be called *structural a priori* information D_S.

Thus we can write for the complete *a priori* information $D_{apriori} = \{D_S, D_0\}$, but typically we will omit D_S from the notation. (Representing the structure of the model itself the structural prior D_S can not be completely represented as a variable within that model.)

Of particular practical interest are independent data measuring a fixed (or 'frozen'), but unknown state ϕ. In that case $x = (x_1, x_2, \cdots)$ and $y = (y_1, y_2, \cdots)$ are vectors with independent components, i.e., $p(y|x,\phi) = \prod_{i=1}^{n} p(y_i|x_i,\phi)$ and $p(x) = \prod_{i=1}^{n} p(x_i)$ factorize. Then

$$p(x,y,\phi|D_0) = p(\phi|D_0) \prod_{i=1}^{n} p(y_i|x_i,\phi)\, p(x_i). \tag{2.5}$$

The x_i and y_i can themselves also be vectors. For Bayesian prediction it turns out to be sufficient to work with the conditional model (see Fig. 2.2),

$$p(y,\phi|x,D_0) = \prod_{i=1}^{n} p(y_i|x_i,\phi)\, p(\phi|D_0). \tag{2.6}$$

However, when evaluating learning algorithms, it is useful to include a sampling model for the x variables and work with (2.5).

We remark that for a given learning situation (2.5) the *a priori* information D_0 on ϕ represents a constant. It is thus not necessary to include D_0 explicitly in the formulae. For example, it is sufficient to write $p(\phi)$ instead of $p(\phi|D_0)$. Nevertheless, we write here D_0 explicitly because its specification for a given learning scenario is an essential part of the model definition. Indeed, nonparametric predictions do in general depend sensitively on D_0.

The interpretation of the model (2.6) is as follows: Preparing an unknown but fixed state ϕ of Nature according to $p(\phi|D_0)$,

6. n independent (tuples of) *training data* $D = \{(x_i, y_i)|1 \leq i \leq n\} = \{(x_D, y_D)\}$ are sampled according to the *likelihood* $p(y_i|x_i,\phi)$ [for given $x_D = \{x_i|1 \leq i \leq n\}$ or for x_D sampled from given $p(x)$].

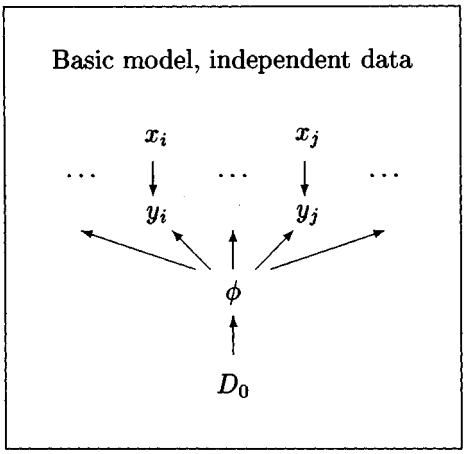


Figure 2.2: Graphical representation of the probabilistic model (2.6) with independent data (conditioned on 'question' x_i and unknown state ϕ), factorizing as follows: $p(y, \phi|x, D_0) = p(\phi|D_0) \prod_{i=1}^{n} p(y_i|x_i, \phi)$. (Here x denotes the vector of all x_i and y the vector of all y_i.) For x sampled according to $p(x) = \prod_i p(x_i)$ this becomes model (2.5) $p(x, y, \phi|D_0) = p(\phi|D_0) \prod_i p(x_i)p(y_i|x_i, \phi)$. To distinguish parametric and nonparametric models we write $p(y|x, \xi)$ for a likelihood parameterized by parameters ξ $(= \phi)$.

The dependent visible variables y_i represent measurement results which are supposed to depend on the visible and hidden causes represented by x_i and ϕ. The visible variables x_i describe known or controlled conditions of measurement. They encode, for example, the kind of measurement device used, the conditions (e.g., time, location) of measurement, and controlled variations of the measured system. Sometimes it is useful to refer to all available (training and prior) data

7. as the (posterior) *state of knowledge* $f = \{D, D_0, (D_S)\}$

Analogously a *prior state of knowledge* $f_0 = \{D_0, (D_S)\}$ can be defined. (Again, we will skip D_S in the following.) A (posterior) state of knowledge f is completely characterized by its *posterior* for the possible state of Nature ϕ,

$$p(\phi|f) = p(\phi|D, D_0), \tag{2.7}$$

a prior state of knowledge f_0 analogously by its prior

$$p(\phi|f_0) = p(\phi|D_0). \tag{2.8}$$

The (posterior) predictive density under f ('likelihood of f')

$$p(y|x, f) = \int d\phi\, p(y|x, \phi)\, p(\phi|f), \tag{2.9}$$

and also the prior predictive density under f_0 ('likelihood of f_0')

$$p(y|x, f_0) = \int d\phi\, p(y|x, \phi)\, p(\phi|f_0), \tag{2.10}$$

are in the convex hull spanned by the likelihoods $p(y|x, \phi)$, i.e., the predictive or data generating densities of the (pure) states of Nature ϕ. (The integral over ϕ and the calculation of the posterior $p(\phi|f)$ will be discussed in Sections 2.1.2 and 2.1.3.) As $p(y|x, f_0)$ is equivalent to $p(y|x, D_0)$ (if skipping D_S) we will not need the notation f_0 in the following. The basic quantities are summarized in Table 2.1, the basic probabilities (or densities, stochastic processes, respectively) in Table 2.2.

Some likelihood models: The introduced setting is quite general. It includes density estimation, classification, and regression problems. In particular, problems with continuous y are called *density estimation* problems, problems with discrete y–variables *classification* problems. (An x–variable is often not included in density estimation.) In (Gaussian) *regression* problems one restricts to Gaussian likelihoods $p(y|x, \phi)$ with fixed variance σ^2 [522],

$$p(y|x, \phi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|y-\phi(x)|^2}{2\sigma^2}}, \tag{2.11}$$

and the function of interest becomes the mean or regression function $\phi(x) = \int dy\, y\, p(y|x, \phi)$. Similarly, in classical *inverse problems*[1] the likelihood is formulated in terms of a ϕ which is considered to be measured only indirectly [222, 262, 478],

$$p(y|x, \phi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|y-(\mathbf{F}\phi)(x)|^2}{2\sigma^2}}. \tag{2.12}$$

Data $D = \{(x_i, y_i)|1 \leq i \leq n\}$ are in such cases sampled for $(\mathbf{F}\phi)(x)$ and not for $\phi(x)$. If the operator $\mathbf{F}$, representing for example an 'instrument or apparatus function', is linear, then $(\mathbf{F}\phi)(x) = \int dx'\, \mathbf{F}(x, x')\phi(x')$.

General non–Gaussian regression problems are obtained if the Gaussian likelihood is replaced by any other fixed–form density with a free and unknown x–dependent mean function $\phi(x)$ to be reconstructed from the data. For counting events, for example, often the choice of a Poisson likelihood (y counts with mean $\phi(x)$ at x)

$$p(y|x, \phi) = [\phi(x)]^y e^{-\phi(x)}/y!, \quad y \in N,\ \phi > 0, \tag{2.13}$$

is appropriate [65, 314, 457], or similarly the choice of a binomial likelihood

$$p(y|x, n; \phi) = \binom{n(x)}{y} [\phi(x)]^y [1 - \phi(x)]^{n(x)-y}, \quad y \in N,\ y \leq n \in N,\ 0 \leq \phi \leq 1, \tag{2.14}$$

[1] Here the term 'inverse problems' is used in the narrower sense referring to the inversion of the usually ill–conditioned operator $\mathbf{F}$. Using the term in a wider sense all empirical learning problems where a general law is inferred from observational data are called inverse problems.

Symbol	Description
x	observable (visible) independent variables ('measurement conditions')
y	observable (visible) dependent variables ('measurement results')
ϕ	hidden variables ('(possible, pure) state of Nature', e.g., 'field') [State defines (is defined by) likelihood $p(y\|x,\phi)$, before learning: prior $p(\phi\|D_0)$, after learning: posterior $p(\phi\|D,D_0) = p(\phi\|f)$]
D_0	(explicit) prior data [defines (are defined by) prior $p(\phi\|D_0)$]
D_S	(implicit) structural prior [most times skipped from the notation]
D	$= (x_D, y_D) = \{(x_i, y_i)\|1 \leq i \leq n\}$, training data, [$y_i$ sampled according to likelihood $p(y\|x,\phi)$, x_i given or x_i sampled from some $p(x)$]
f	$= \{D, D_0\}$, (posterior) state of knowledge [State of knowledge defines (is defined by) posterior $p(\phi\|f)$, determines (posterior) predictive density $p(y\|x,f) = \int d\phi\, p(y\|x,\phi)p(\phi\|f)$]
f_0	$= \{D_0\}$, prior state of knowledge [State of knowledge defines (is defined by) prior $p(\phi\|f_0)$, determines prior predictive density $p(y\|x,f_0) = \int d\phi\, p(y\|x,\phi)p(\phi\|f_0)$]
a	actions (possible decisions, e.g., approximations, see Section 2.2)
$l(x,y,a)$	loss function (for outcome y under condition x and action a, see Section 2.2)

Table 2.1: The basic quantities.

Probability	Description
$p(\phi\|D_0)$	prior
$p(\phi\|f) = p(\phi\|D, D_0)$	posterior
$p(y\|x, \phi)$	likelihood, predictive density for a pure state
$p(y\|x, D_0)$	prior predictive density
$p(y\|x, f) = p(y\|x, D, D_0)$	(posterior) predictive density

Table 2.2: The basic probabilities (or densities, stochastic processes, respectively).

if at x from $n(x)$ binary events a number of y 'ones' and $n(x) - y$ 'zeros' have been observed, $\phi(x)$ being the probability of 'one' at x.

A classification problem, for instance, is face detection, where the input x is a highdimensional vector (a graylevel image) and outcome y is simply zero or one, depending on whether the image shows a face or not. The unknown state ϕ represents the optimal face recognizer to be approximated. In image reconstruction, pixel values y_i are measured at coordinates x_i from an unknown original image ϕ. In its most general form this represents a density estimation problem, while under the assumption of Gaussian noise this becomes a regression problem. Similarly, in reconstructing noisy time series, the y_i can denote values obtained at times x_i measuring a true but hidden time development ϕ.

Technically, the variables ϕ, which are unknown but assumed fixed, determine the likelihood $p(y|x, \phi)$. In this book we will mainly be interested in *nonparametric* models where ϕ represents a function of x and/or y. In particular we will be interested in nonparametric density estimation where the (x, y)–dependent numbers $p(y|x, \phi)$ are considered as the primary degrees of freedom which 'parameterize' a model state ϕ. Similarly, we will speak of nonparametric regression if each value $\phi(x)$ of the regression function is considered as a single degree of freedom (see Eq. (2.11)). If the number of parameters (i.e., the dimension of ϕ) is finite and small compared to the possible degrees of freedom of the likelihood, we will call the model *parametric*. If we want to refer explicitly to a parametric model we will write ξ for ϕ. A typical example of a parametric model with a two–dimensional parameter vector $\xi = \xi_0, \xi_1$ is obtained if one assumes a Gaussian likelihood with a fixed constant

variance and a regression function restricted to a linear function $\xi_0 + \xi_1 x$

$$p(y|x,\xi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|y-\xi_0-\xi_1 x|^2}{2\sigma^2}}. \tag{2.15}$$

Analogously, using a polynomial $\sum_{k=0}^{q} \xi_k x^k$ of degree q for the regression function gives

$$p(y|x,\xi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|y-\sum_{k=0}^{q} \xi_q x^q|^2}{2\sigma^2}}. \tag{2.16}$$

In the case of a general nondiagonal data covariance matrix $C_{ij} = K_{ij}^{-1}$ Eq. (2.16) becomes

$$p(y_1,\cdots,y_n|x_1,\cdots,x_n,\xi) =$$

$$\left(\det \frac{K}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\sum_{i,j}(y_i-\sum_{k=1}^{q} \xi_q x_i^q)K_{ij}(y_j-\sum_{k=1}^{q} \xi_q x_j^q)}. \tag{2.17}$$

Similarly, replacing $y - \sum_{k=1}^{q} \xi_q x^q$ by some arbitrary fixed function $F(x,y,\xi)$, leads to

$$p(y|x,\xi) = \left(\det \frac{K}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\sum_{i,j} F(x_i,y_i,\xi)K_{i,j}F(x_j,y_j,\xi)}. \tag{2.18}$$

Models like in Eq. (2.18) are used in the generalized method of moments (GMM) [18,78,97,197,434].

As one can consider parametric models in the limit where the number of parameters goes to infinity, and as nonparametric models have to be discretized for numerical calculations, the distinction between parametric and nonparametric models is not sharp in practice. Fig. 2.3 displays nonparametric and parametric approaches graphically.

Some specific noise models: The basic model (2.3) can be modified in many ways, some of them are shown in Fig. 2.4. For instance, including *input noise*,

$$p(y|x,\phi) = \int d\tilde{x} p(y|\tilde{x},\phi)\, p(\tilde{x}|x), \tag{2.19}$$

means that the value $\tilde{x}$ for which y is observed can only be controlled stochastically. Similarly, *output noise*,

$$p(y|x,\phi) = \int d\tilde{y} p(y|\tilde{y})\, p(\tilde{y}|x,\phi), \tag{2.20}$$

models additional noise processes when reading off measurement results. Clearly, if input and output noise are combined, the (hidden) likelihood $p(\tilde{y}|\tilde{x},\phi)$ has to be integrated over $\tilde{x}$ and $\tilde{y}$.

In Section 2.3.2 we will study how priors for the effective (visible) likelihoods $p(y|x,\phi)$ can be related to priors for (hidden) likelihoods $p(\tilde{y}|\tilde{x},\phi)$.

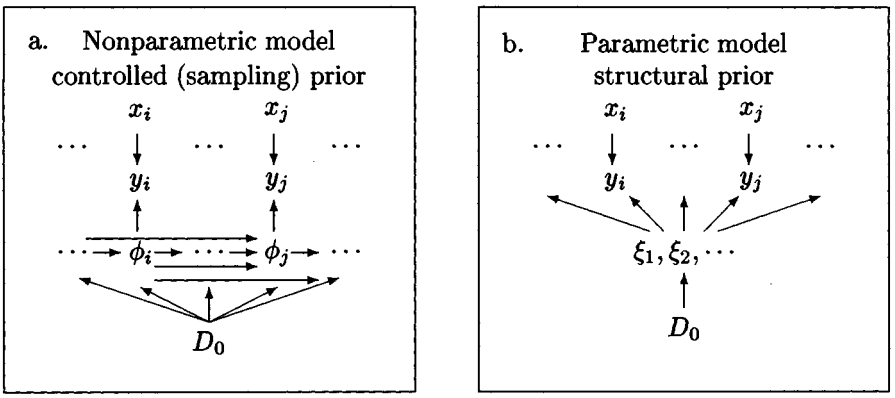


Figure 2.3: Schematic representation of nonparametric and parametric models. a. Nonparametric model: Shown is an example where the hidden variables $\phi(x, y)$ directly represent the values of the likelihood, $p(y_i|x_i, \phi) = \phi(x_i, y_i) = \phi_i$. In that case ϕ has to be non–negative and normalized over y. The unknown state ϕ (with components ϕ_i) is sampled according to $p(\phi|D_0) = p(\phi_1|D_0)p(\phi_2|\phi_1, D_0)p(\phi_3|\phi_1, \phi_2, D_0)\cdots$ choosing some ordering ('time') of the ϕ_i. (For example, smooth regression functions $\phi(x)$ can be sampled from a Gaussian process with smoothness–related covariance.) Because the dependencies between data are determined by sampling ϕ under the control of D_0 this is called a *controlled (sampling) prior* (see Section 2.3.1). b. Parametric model: $p(y, \xi|x, D_0) = p(\xi|D_0) \prod_i p(y_i|x_i, \xi)$ for parameter vector ξ with components ξ_k. Here the different likelihood values $p(y|x, \xi)$ are dependent because they are determined by the same parameters ξ. These structural dependencies represent a structural prior D_S. (See also Section 2.3.1.) To compare b. with a., one can introduce the individual likelihood values $\phi_i(\xi) = \phi(x_i, y_i; \xi) = p(y_i|x_i, \xi) = p(y_i|x_i, \phi(\xi))$ as variables which depend deterministically on ξ, i.e., $p(\phi_i|\xi) = \delta(\phi_i - \phi_i(\xi))$. The model including ϕ_i reads $p(y, \xi, \phi|x, D_0) = p(y, \xi|x, D_0) = p(\xi|D_0) \prod_i p(y_i|x_i, \phi_i(\xi))$. For a probabilistic relation $p(\phi_i|\xi)$ the parameter ξ is called *hyperparameter* and will be written θ to distinguish it from 'deterministic' parameters ξ. (See Fig. 2.4.) Technical remark: Fig. 2.3 gives a graphical representation slightly different from Fig. 2.2. That part of D_0 in Fig. 2.2 which contains the dependencies of the ϕ_i is made explicit in Fig. 2.3a. Furthermore, note that the probability density $p(y_i|x_i, \phi)$ considered as function of varying x_i, y_i does depend on the whole function ϕ. This is shown in Fig. 2.2. In contrast Fig. 2.3a. illustrates that the probability density $p(y_i|x_i, \phi)$ for fixed x_i, y_i depends only on the function value $\phi(x_i, y_i)$. In particular, if $x_j = x_i$ and $y_j = y_i$ then $\phi_i = \phi_j$ stands for only one random variable with one fixed value. In case all x_i are different this distinction can be ignored.

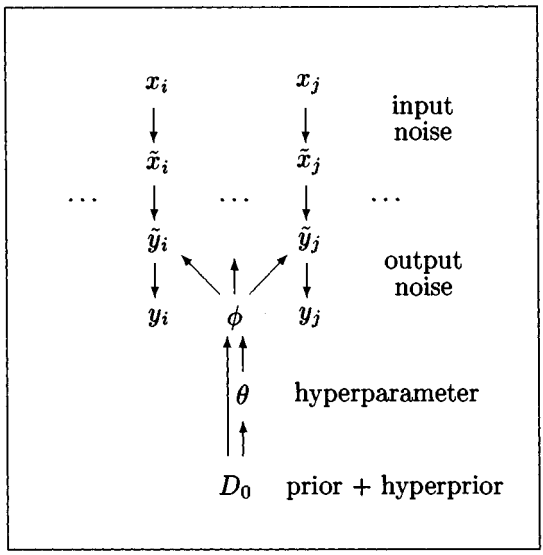


Figure 2.4: Additional noise processes (input noise, output noise, and hyperparameters introducing additional noise in the preparation of ϕ). $p(y, \phi|x, D_0) = \int d\tilde{x}\, d\tilde{y}\, d\theta\, p(\theta|D_0)p(\phi|\theta, D_0) \prod_i p(y_i|\tilde{y}_i)p(\tilde{y}_i|\tilde{x}_i, \phi)p(\tilde{x}_i|x_i)$. The integral $\int d\tilde{x}\, d\tilde{y}\, d\theta$ over the vectors $\tilde{x}$, $\tilde{y}$, θ is understood as integration over all components $\tilde{x}_i$, $\tilde{y}_i$, θ_k.

Indeed, according to the specific structure of the effective likelihoods in (2.19) and (2.20), unspecific nonparametric priors for the hidden likelihoods can translate into more informative priors for the effective likelihoods. (For example, output noise with $p(y|\tilde{y}) = p(y)$ leaves the effective likelihood of Eq. (2.20) completely independent of ϕ.) In particular we will see how noise processes can lead to general smoothness priors.

Similarly, random variables θ can introduce additional *preparation noise* for sampling the state ϕ. Such hyperparameters θ parameterize the prior $p(\phi|\theta, D_0)$ [33, 73, 150, 306]. Introducing a so–called hyperprior $p(\theta|D_0)$ and integrating over hyperparameters leads to the prior

$$p(\phi|D_0) = \int d\theta p(\phi|\theta, D_0)p(\theta|D_0). \tag{2.21}$$

Hyperparameters are useful to construct complex priors $p(\phi|D_0)$ from simple building blocks $p(\phi|\theta, D_0)$. For example, consider Gaussian process priors $p(\phi|\theta, D_0)$ with mean and covariance parameterized by θ. For a hyperparameter θ taking a finite number of discrete values the resulting prior $p(\phi|D_0)$ is a finite mixture of Gaussian processes.

Empirical learning: Let us understand by *empirical learning* the *prediction of data (test data)* based on (a finite number of) *observational data (training data)*. Clearly, to be able to make predictions there must exist dependencies between test and training data. Knowledge of these dependencies is essential for learning and represents the *a priori* information. The probabilistic model has to implement the available *a priori* information for the problem under study. Hence successful learning requires both, observational data D and adequate *a priori* information D_0 (and D_S). We will see that observational data D are used to update the prior $p(\phi|D_0)$ to obtain the *a posteriori* density or *posterior* $p(\phi|f) = p(\phi|D, D_0)$ where $f = \{D, D_0, D_S\}$ stands for the *posterior state of knowledge*. [We write $p(\phi|D, D_0)$ for the posterior if we want to emphasize the dependency on data D and prior D_0 (skipping D_S), but use $p(\phi|f)$ in general contexts.] Having calculated the posterior one is able to make predictions using the (posterior) predictive density (2.9), i.e., $p(y|x, f) = p(y|x, D, D_0) = \int d\phi\, p(y|x, \phi) p(\phi|D, D_0)$. Thus the basic steps of Bayesian learning to be discussed in the following are

1. updating of the *a priori* distributions for the hidden variables based on the observational data to obtain new (*a posteriori*) distributions, and

2. eliminating the hidden variables by summing (marginalizing) over them to obtain a prediction.

Finally, let us remark that the framework of probabilistic models is now widely used for probabilistic expert systems, known in this context as *Bayesian networks*, *causal networks*, or *belief networks* [241,245,282,386,416]. They allow to model relatively complex structures. Typically they restrict to discrete variables and parametric approaches. Like for example in Fig. 2.4 they can conveniently be represented in form of a directed acyclic graph, where the arrows are often, but not necessarily, interpreted as a causal structure. To represent an independence structure it is sufficient to use undirected graphs (without arrows). The corresponding factorization properties of a joint probability density can be expressed as a *Markov random field* or, equivalently, as a *Gibbs random field* [41,83,194]. For updating probabilistic models under new data there exists a general 'local message passing' or 'junction tree' algorithm [96,242,447,451]. In Section 2.1.2 probabilistic models will be expressed using notations of statistical physics.

2.1.2 Energies, free energies, and errors

Before we discuss the Bayesian approach to empirical learning we introduce in this section some notations connecting probabilistic models to statistical physics. This allows to use for empirical learning the language and methods of statistical physics (or of Euclidean field theory).

Log–probability and independent events: It is often more convenient to work with *log–probabilities*

$$L = \ln p \tag{2.22}$$

than with probabilities p (or with log–densities instead of densities, or with log–processes instead of stochastic processes, respectively):

1. This ensures non–negativity of $p = e^L \geq 0$ for arbitrary L. Thus when working with log–probabilities one can skip the non–negativity constraint, necessary when working with probabilities.
2. The multiplication of probabilities for independent events, yielding their joint probability, becomes a sum when written in terms of L.

Indeed, from

$$p(A, B) = p(A \text{ AND } B) = p(A)p(B) \tag{2.23}$$

it follows for

$$\begin{aligned} L(A, B) &= L(A \text{ AND } B) = \ln p(A, B) \\ &= \ln(p(A)p(B)) = \ln p(A) + \ln p(B) \\ &= L(A) + L(B). \end{aligned} \tag{2.24}$$

Especially in the limit where an infinite number of events is combined by AND, this results in an infinite product for p but yields an integral for L, which is easier to treat. Similarly, $b = -L = -\ln p$ is known as *bit number* or *(self) information* (see Appendix B.2.2), its expectation value is called *average (self) information* (see Appendix B.3). For example, a Gaussian likelihood as in (2.11) has the log–probability

$$L(y|x, \phi) = \ln p(y|x, \phi) = -\frac{|y - \phi(x)|^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2), \tag{2.25}$$

and the product $\prod_{i=1}^n p(y_i|x_i, \phi)$ becomes a sum of log–likelihoods

$$\begin{aligned} L(y|x, \phi) &= \ln \prod_{i=1}^n p(y_i|x_i, \phi) = \sum_{i=1}^n \ln p(y_i|x_i, \phi) \\ &= -\sum_{i=1}^n \frac{|y_i - \phi(x_i)|^2}{2\sigma^2} - \frac{n}{2}\ln(2\pi\sigma^2). \end{aligned} \tag{2.26}$$

Defining the empirical densities

$$p_{\text{emp}}(x, y) = \frac{1}{n}\sum_{i=1}^n \delta(x - x_i)\delta(y - y_i), \tag{2.27}$$

$$p_{\text{emp}}(x) = \frac{1}{n}\sum_{i=1}^n \delta(x - x_i), \tag{2.28}$$

$$p_{\text{emp}}(y|x_i) = \frac{p_{\text{emp}}(x_i, y)}{p_{\text{emp}}(x_i)} \tag{2.29}$$

with $x_i \in D$ so that $p_{\text{emp}}(y|x_i)$ is defined, we can also write

$$\begin{aligned} p(y|x,\phi) &= \prod_{i=1}^{n} p(y_i|x_i,\phi) \\ &= e^{\sum_{i=1}^{n} \ln p(y_i|x_i,\phi)} \\ &= e^{n \int dx'\, dy'\, p_{\text{emp}}(x',y') \ln p(y'|x',\phi)} \\ &= e^{n \int dx'\, p_{\text{emp}}(x') \int dy'\, p_{\text{emp}}(y'|x') \ln p(y'|x',\phi)} \\ &= e^{-n \int dx'\, p_{\text{emp}}(x') K(p_{\text{emp}}(y'|x'),p(y'|x',\phi))}, \end{aligned} \tag{2.30}$$

i.e.,

$$\begin{aligned} L(y|x,\phi) &= \sum_{i=1}^{n} \ln p(y_i|x_i,\phi) \\ &= -n \int dx'\, p_{\text{emp}}(x') K(p_{\text{emp}}(y'|x'),p(y'|x',\phi)) \\ &= -nK_c(p_{\text{emp}}(x',y'),p(y'|x',\phi)), \end{aligned} \tag{2.31}$$

or in short $L = \sum_i L_i = -nK_c$, where

$$K(p_{\text{emp}}(y'|x'),p(y'|x',\phi)) = -\int dy'\, p_{\text{emp}}(y'|x') \ln p(y'|x',\phi) \tag{2.32}$$

is the *Kerridge inaccuracy* [256] of the likelihood $p(y|x,\phi)$ with respect to the conditional empirical density $p_{\text{emp}}(y|x)$ and

$$\begin{aligned} K_c(p_{\text{emp}}(x',y'),p(y'|x',\phi)) &= -\int dx'\, p_{\text{emp}}(x') K(p_{\text{emp}}(y'|x'),p(y'|x',\phi)) \\ &= -\int dx'\, dy'\, p_{\text{emp}}(x',y') \ln p(y'|x',\phi) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \ln p(y_i|x_i,\phi) \end{aligned} \tag{2.33}$$

the corresponding *conditional Kerridge inaccuracy.* We may also remark that, written this time for discrete x, the Kerridge inaccuracy

$$K(p,q) = -\sum_x p(x) \ln q(x) \tag{2.34}$$

is the sum of the *entropy* [448], sometimes also called 'measure of uncertainty',

$$H(p) = -\sum_x p(x) \ln p(x) \tag{2.35}$$

and the *Kullback–Leibler entropy* or *Kullback–Leibler distance* [276]

$$KL(p,q) = -\sum_x p(x) \ln \frac{p(x)}{q(x)}, \tag{2.36}$$

i.e.,

$$K(p,q) = H(p) + KL(p,q). \tag{2.37}$$

Partition sums and normalization: Besides the requirement of being non–negative, probabilities have to be normalized, $\int dx\, p(x) = 1$. (Here x can stand for any random variable. For the sake of simplicity we use the integral notation, and probability can also stand for a density or a stochastic process.) To work with *unnormalized probabilities* $Z(x) \geq 0$, writing

$$p(x) = \frac{Z(x)}{Z} \tag{2.38}$$

with norming constant or *partition sum* (for a slightly more general definition of partition sums see Appendix B.2.1)

$$Z = \int dx\, Z(x), \tag{2.39}$$

can thus have advantages:

1. With this paramaterization of probability the normalization is always explicit.

2. When dealing with a large set of elementary events normalization is numerically a nontrivial task. It is then convenient to work as far as possible with unnormalized probabilities $Z(x)$. For example, for maximizing a probability $p(x)$ with respect to x no normalization is needed.

For the example of a Gaussian likelihood (2.11) the partition sum or norming constant is $Z = \sigma\sqrt{2\pi}$. Clearly, to find the maximum of the likelihood (2.11) at $\phi(x) = y$ the norming constant does not have to be calculated, because Z is independent of ϕ.

Energy, free energy, and temperature: Applying the logarithmic transformation to unnormalized probabilities the advantages can be combined,

$$g(x) = \ln Z(x). \tag{2.40}$$

Furthermore it is for technical reasons often useful to introduce an extra factor β, yielding

$$E(x) = -(1/\beta) \ln Z(x). \tag{2.41}$$

In statistical physics, $E(x)$ is known as *energy*, $F = -(1/\beta) \ln Z$ as *free energy*, and β is (proportional to) the *inverse temperature*. (For a slightly more general definition of energy, see Appendix B.3; for a slightly more general definition of free energy, B.2.2; and for temperature, B.4). Hence we will in the following often write

$$p(x) = \frac{e^{-\beta E(x)}}{Z}, \tag{2.42}$$

or in words,

$$\text{probability} = \frac{1}{\text{partition sum}} \exp\left(-\frac{\text{energy}}{\text{temperature}}\right). \tag{2.43}$$

As one always has a free constant when choosing an energy for p, so, for example, $E = -L$ is also a possible choice for which $Z = 1$. Note that changing the constant changes the definition of the temperature. For the example of a Gaussian likelihood (2.11) an energy is, if choosing $\beta = 1$, up to an arbitrary constant

$$E(y|x, \phi) = \frac{|y - \phi(x)|^2}{2\sigma}, \tag{2.44}$$

or, if we identify β with $1/\sigma^2$,

$$E(y|x, \phi) = \frac{1}{2}|y - \phi(x)|^2. \tag{2.45}$$

Varying the inverse temperature β generates an exponential family of densities [13, 26, 282, 438]. An (inverse) temperature can be utilized for several purposes:

1. A temperature parameter can be utilized to control the number of local maxima when a sum of probabilities is considered. This is frequently used in practice by (simulated or deterministic) annealing techniques which minimize free energies [1, 98, 149, 261, 338, 415, 428, 532–534]. (See also Appendix B.4.2.) As mentioned above, for a Gaussian likelihood (2.11) the (likelihood) temperature can be identified with the variance σ^2. Changing temperature thus changes the width of the Gaussian, the low temperature limit ($\sigma^2 \to 0$) approaching a δ-peak, the high temperature limit ($\sigma^2 \to \infty$) approaching a flat distribution. Accordingly, a mixture of two Gaussians will have two local maxima at low temperatures and only one local maximum at high temperatures. As changing the temperature can thus change the number of local maxima we will use these limits for example in Sections 6.4.2 and 6.6 for mixture models.

2. A temperature parameter allows to interpret a probability as a specific maximum entropy solution. In physics the inverse temperature β plays the role of a Lagrange multiplier in the maximum entropy approach to statistical physics [23, 67]. Indeed, every probability density $p(x)$ can

be interpreted as least informative density ((Boltzmann–)Gibbs distribution, generalized canonical ensemble) obtained by maximizing its *entropy* (2.35), i.e.,

$$H(p) = -\sum_x p(x) \ln p(x) \tag{2.46}$$

(cf. also the slightly more general definition of entropy in Appendix B.3.2) under the constraints $\sum_x p(x) = 1$ and $\sum_x E(x)p(x) = U$, the Lagrange parameter β being determined by the value of U [23, 480]. (For details see Appendix B.4.1.) For continuous variables technical difficulties are avoided by working with the *relative entropy* or *Kullback–Leibler entropy* (cf. Eq.(2.36)) for densities

$$KL(p, p_0) = -\int dx\, p(x) \ln \frac{p(x)}{p_0(x)}, \tag{2.47}$$

with p_0 representing a reference density, which leads to $p \propto p_0 e^{-\beta E}$ [23, 458].

3. A temperature–like parameter makes it possible to calculate averages (moments and cumulants) by taking derivatives using the method of generating functions. Indeed, the free energy F is a generating function for the cumulants of the energy, which means that cumulants of E can be obtained by taking derivatives of F with respect to β [23, 29, 142, 351]. For example, the average of the random variable $E(x)$ under $p(x) = e^{-E(x)}/Z$ can be calculated from the free energy F as follows

$$\begin{aligned} -\beta \frac{\partial}{\partial \beta} F &= -\frac{\partial}{\partial \beta} \ln Z \\ &= -\frac{\partial}{\partial \beta} \ln \int dx\, e^{-\beta E(x)} \\ &= \frac{\int dx\, E(x) e^{-\beta E(x)}}{\int dx\, e^{-\beta E(x)}} \\ &= \int dx\, E(x) p(x) = \langle E \rangle . \end{aligned} \tag{2.48}$$

In the language of physics the inverse temperature β can be seen as an external field coupling to the energy. (The interested reader may look in Appendix B.4.3 for details.)

When not needed as explicit parameter, the inverse temperature β is set to one.

In the following we will mainly need log–probabilities, unnormalized probabilities, and energies. A more detailed discussion of the relations between probability, log–probability, energy, bit numbers, information, free energy,

partition sums, temperature, moment and cumulant generating functions, and effective actions can be found in Appendix B (and in [294]).

Application to likelihood, prior and posterior: Since a probability can be expressed in various different forms, it is convenient to use the same notation for all the variants, as we already did for example in Eqs. (2.25, 2.44, 2.45). For example, for the likelihood $p(y|x,\phi)$ we write for a corresponding unnormalized probability $Z(y|x,\phi)$, a corresponding (likelihood) energy $E(y|x,\phi)$, and for a (likelihood) free energy $F(Y|x,\phi) = -\ln Z(Y|x,\phi)/\beta'$, where Y indicates integration over $y \in Y$ and β' denotes an inverse temperature for the likelihood. Thus we write for likelihoods

$$p(y|x,\phi) = e^{-\beta'(E(y|x,\phi)-F(Y|x,\phi))}. \tag{2.49}$$

Analogously, the posterior $p(\phi|D,D_0) = p(\phi|f)$ in state of knowledge $f = f(D,D_0)$ can be written as

$$p(\phi|f) = e^{L(\phi|f)} = \frac{Z(\phi|f)}{Z(\Phi|f)} = \frac{e^{-\beta E(\phi|f)}}{Z(\Phi|f)} = e^{-\beta(E(\phi|f)-F(\Phi|f))}, \tag{2.50}$$

with $\phi \in \Phi$ and *log-posterior*

$$L(\phi|f) = \ln p(\phi|f), \tag{2.51}$$

unnormalized posterior

$$Z(\phi|f) = e^{-\beta E(\phi|f)}, \tag{2.52}$$

posterior partition sum

$$Z(\Phi|f) = \int_{\phi\in\Phi} d\phi\, Z(\phi|f) = \int_{\phi\in\Phi} d\phi\, e^{-\beta E(\phi|f)}, \tag{2.53}$$

posterior energy

$$E(\phi|f) = -\frac{1}{\beta}\ln Z(\phi|f), \tag{2.54}$$

for *posterior temperature* $1/\beta$, and *posterior free energy*

$$F(\Phi|f) = -\frac{1}{\beta}\ln Z(\Phi|f) = -\frac{1}{\beta}\ln \int_{\phi\in\Phi} d\phi\, e^{-\beta E(\phi|f)}. \tag{2.55}$$

In the parametric case $\int d\phi$ represents an integral over parameters ξ. In the nonparametric case $\int d\phi$ stands for a functional (or path) integral over functions $\phi(x,y) = p(y|x,\phi)$ in density estimation or $\phi(x)$ in regression. Thus we see that the posterior energy (and also the prior energy, see below) can be interpreted as the Lagrangian of a *statistical field theory* for Euclidean scalar fields ϕ [232,284,541].

The same conventions will also be used for other probabilities (densities, stochastic processes), so we write for the prior

$$p(\phi|D_0) = e^{-\beta''(E(\phi|D_0)-F(\Phi|D_0))}. \tag{2.56}$$

Two of the inverse temperatures β, β', β'' for posterior, likelihood, and prior can be chosen independently. For the sake of simplicity we did not include the temperature–dependency of the functions Z, F in our notation.

Error and energy: In Section 2.1.6 we will discuss the Maximum A Posteriori Approximation where an optimal ϕ is found by maximizing the posterior $p(\phi|f)$. Since maximizing the posterior means minimizing the posterior energy $E(\phi|f)$, the latter plays the role of an *error functional* for ϕ to be minimized. This is technically similar to the minimization of a regularized error functional as it appears in regularization theory or empirical risk minimization, discussed in Section 2.2.5.

Functional integrals: Let us finally have a closer look at the integral over model states ϕ in the nonparametric case. Interpreted as random variable ϕ corresponds to a stochastic process. In the case of regression the resulting integral is a functional integral over real functions $\phi(x)$

$$\int d\phi \to \int \left(\prod_x d\phi(x) \right). \tag{2.57}$$

Similarly, nonparametric density estimation requires integration over functions $\phi(x, y)$, but with additional non–negativity and normalization constraint on $\phi(x, y) = p(y|x, \phi)$, which can be implemented by using δ-functions and step functions (see Section 2.1.7). Gaussian functional integrals as they appear in regression problems can be rigorously defined [111,165,309,322,507]. However, a strict definition of non–Gaussian functional integrals, as they appear in density estimation, leads to mathematical difficulties which are highly non–trivial. (For perturbative approaches to the construction of continuum limits for non–Gaussian processes see for instance [72,84,107,232,397,435,515,516,541], for a non–perturbative ϕ^4–theory see [165], for renormalization in the context of density estimation see [6].) In this book we will take in these cases the numerical point of view where all functions are considered to be finally discretized and the ϕ–integral is well–defined (cf. for example with the 'lattice regularization' approach [91,351,430]). Clearly, there is always a huge freedom in the choice of the basis in which a theory is discretized.

2.1.3 Bayes' theorem: Posterior, prior, and likelihood

Learning means updating propositions about unknown states of Nature on the basis of observational data. Technically, this requires the calculation of

posterior densities $p(\phi|D, D_0)$ from prior $p(\phi|D_0)$ and likelihood $p(y|x, \phi)$. Posteriors are linked to likelihoods by Bayes' theorem

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}, \tag{2.58}$$

which follows at once from the definition of conditional probabilities, $p(H, D) = p(H|D)p(D) = p(D|H)p(H)$ as $P(D) \neq 0$ for observed data. In a Bayesian context H represents the hidden variables (hypotheses, theories, models, in nonparametric approaches described by fields ϕ) and D the available data. The terms in Eq. (2.58) are referred to as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}, \tag{2.59}$$

where the evidence $p(D)$ is also known as marginalized density of the data.

Bayes' theorem (2.58) can be interpreted in two ways. First, one can speak of inversion of $p(D|H)$ to get $p(H|D)$, necessary to select an optimal H. This is the reason why the 'posterior' $p(H|D)$ has been called 'inverse probability' by Laplace. Second, one can emphasize that the 'prior' $p(H)$ is updated under the new information D to get the posterior $p(H|D)$. Both interpretations are useful. A simple application of Bayes' theorem is illustrated in Section 2.1.8.

The application of Bayes' theorem to data and hypotheses has been heavily discussed in the past (partly continuing to the present). Indeed, the classical *Frequentist* approach to statistics tries to base the (unobservable) mathematical concept of probability entirely on observed relative frequencies [132, 133]. From this point of view prior probabilities (densities, processes) for hidden variables and, similarly, posterior or 'inverse' probabilities (densities, processes), are not considered as well defined concepts. In particular, hypotheses do not typically represent random variables which can be sampled easily. For instance, if the hypotheses represent different values for the 'true radius of the earth' to be determined by noisy measurements, a (prior or posterior) sample with many different values for the 'true radius of the earth' would be difficult to obtain in practice. Thus Frequentists base statistics on the likelihood or direct probability (density) $p(D|H)$ alone. This restriction is the origin of principal difficulties, for example, if a decision for an optimal action (or approximation) is required, which does not appear within a Bayesian framework. In particular, inference based only on frequencies is not possible in nonparametric settings where typically at most one data point is available for a specific condition, e.g., a specific value of x, and for most conditions no data point is available. Bayesians, on the other hand, being in the tradition of Bayes and Laplace, typically found statistics directly on decision theory. This allows to interpret probability consistently as degree of belief or plausibility for which the same mathematical rules can be used as for mathematical probability. (See for example [33, 87, 88, 102, 175, 234, 238, 257, 311, 314, 386, 446], the references therein, and also Sections 2.2.5 and 2.3.)

According to Bayes' theorem the posterior of ϕ for n independent training data $D = (x_D, y_D) = \{(x_i, y_i) | 1 \leq i \leq n\}$ can be written as (see Fig. 2.5)

$$p(\phi|f) = p(\phi|D, D_0) = \frac{p(D|\phi)\, p(\phi|D_0)}{p(D|D_0)} = \frac{p(y_D|x_D, \phi)\, p(\phi|D_0)}{p(y_D|x_D, D_0)}$$

$$= \frac{p(\phi|D_0) \prod_i p(x_i, y_i|\phi)}{\int d\phi'\, p(\phi'|D_0) \prod_i p(x_i, y_i|\phi')} = \frac{p(\phi|D_0) \prod_i p(y_i|x_i, \phi)}{\int d\phi'\, p(\phi'|D_0) \prod_i p(y_i|x_i, \phi')}, \qquad (2.60)$$

assuming $p(y_D|x_D, \phi, D_0) = p(y_D|x_D, \phi)$ and $p(\phi|D_0, x_D) = p(\phi|D_0)$. Eq. (2.60) shows that the posterior can be expressed equivalently by joint likelihoods $p(y_i, x_i|\phi)$ or by conditional likelihoods $p(y_i|x_i, \phi)$. When working with joint likelihoods, a distinction between y– and x–variables is not necessary. In that case x can be included in y and skipped from the notation. If, however, $p(x)$ is already known or not of interest, working with conditional likelihoods is preferable.

Thus the Bayesian approach to learning requires

1. the specification of a *likelihood model*, i.e., of a space Φ of possible hidden states ϕ, defined by the corresponding likelihoods $p(y|x, \phi)$, and

2. the specification of a *prior model* $p(\phi|D_0)$ for ϕ.

Bayes' theorem (2.60) is then used to obtain a new posterior from a given prior if new data D arrive. The likelihood model $p(y|x, \phi)$ is also sometimes referred to as *observation equation*, and the prior $p(\phi|D_0)$ is related to what is also called the *system equations*. Indeed, decomposing ϕ into components ϕ_i, the prior model can be factorized as $p(\phi|D_0) = p(\phi_1|D_0)\; p(\phi_2|\phi_1, D_0)\; p(\phi_3|\phi_1, \phi_2, D_0) \ldots$, where the single factors represent the probabilistic system equations for ϕ_i. For example, in models where ϕ_{i+1} only depends on ϕ_i the system equations are given by the set of $p(\phi_{i+1}|\phi_i, D_0)$ [517].

In terms of energies Eq. (2.60) reads

$$p(\phi|f) = \frac{e^{-\beta \sum_i E(y_i|x_i, \phi) - \beta E(\phi|D_0)}}{Z(Y_D|x_D, \phi)} \left(\int d\phi' \, \frac{e^{-\beta \sum_i E(y_i|x_i, \phi') - \beta E(\phi'|D_0)}}{Z(Y_D|x_D, \phi')} \right)^{-1}, \qquad (2.61)$$

when choosing the same temperature $1/\beta$ for both energies. The normalization constant of the total likelihood is $Z(Y_D|x_D, \phi) = \prod_i \int dy_i\, e^{-\beta E(y_i|x_i, \phi)}$.

We remark that for continuous x and/or y the likelihood $p(y|x, \phi)$ describes an ideal measurement because realistic measurements can not be arbitrarily sharp. Considering the function $p(\cdot|\cdot, \phi)$ as element of a Hilbert space, its value at x, y can be written as scalar product $p(y|x, \phi) = \langle x, y|\, p(\cdot|\cdot, \phi)\rangle$ with some functional $\langle x, y|$. For continuous x and/or y this notation is only formal as $\langle x, y|$ becomes unnormalizable and is thus not an element in the dual Hilbert space. In practice a measurement of $p(\cdot|\cdot, \phi)$ corresponds to a normalizable

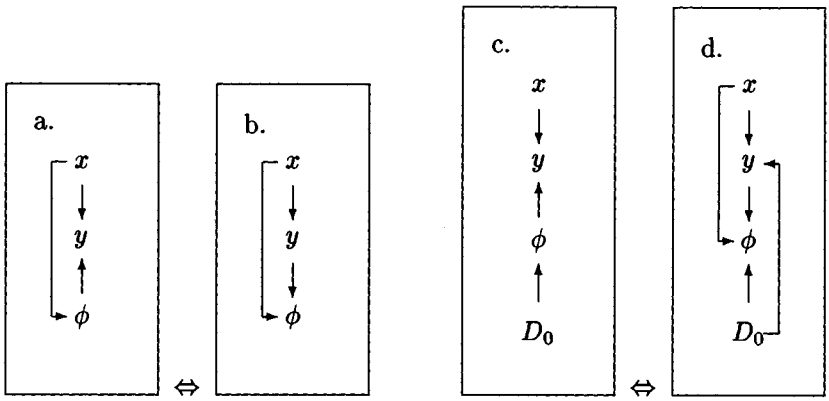


Figure 2.5: Graphical representation of Bayes' theorem (2.60) applied to the likelihood $p(y|x,\phi)$. a., b.: In model (2.2), $p(y|x,\phi)p(\phi|x)p(x) = p(\phi|x,y)p(y|x)p(x)$ with $p(y|x) = \int d\phi\, p(y|x,\phi)p(\phi|x)$. c., d.: In model (2.3), $p(y|x,\phi)p(\phi|D_0)p(x) = p(\phi|x,y,D_0)p(y|x,D_0)p(x)$ with $p(y|x,D_0) = \int d\phi\, p(y|x,\phi)p(\phi|D_0)$.

$\langle\tilde{x},\tilde{y}| = \int dy \int dx\, \vartheta(\tilde{x},\tilde{y};x,y)\,\langle x,y|$.[2] The data terms can then be understood as

$$p(\tilde{y}|\tilde{x},\phi) = \frac{\int dy\, dx\, \vartheta(\tilde{x},\tilde{y};x,y)p(y,x|\phi)}{\int d\tilde{y}\, dy\, dx\, \vartheta(\tilde{x},\tilde{y};x,y)p(y,x|\phi)}. \tag{2.62}$$

The notation $p(y|x,\phi)$ is understood as limit $\vartheta(\tilde{x},\tilde{y};x,y) \to \delta(\tilde{x}-x)\delta(\tilde{y}-y)$ and means in practice that $\vartheta(\tilde{x},\tilde{y};x,y)$ is very sharply centered. We will assume that the discretization, finally necessary in numerical calculations, implements such an averaging.

2.1.4 Predictive density and learning

The aim of learning consists of predicting outcome y in new test situations x. For known state ϕ the probability density of y is given by the likelihood $p(y|x,\phi)$. For unknown state ϕ one has to average the 'pure' likelihoods $p(y|x,\phi)$ over the distribution of the ϕ. Before receiving data this results in

[2]Choosing normalizable functions $\langle\tilde{x},\tilde{y}|$ as coordinates defines a reproducing kernel Hilbert space (RKHS) [19,208,259,260,309,383,507]. As separable Hilbert spaces of the same dimension (e.g., countable) are isomorphic [412], a RKHS defined on a compact domain is, for example, isomorphic to the space of square–integrable functions L^2 on that domain. A Hilbert space isomorphism is given by the square root $L_K^{1/2}$, i.e., the only linear operator satisfying $L_K^{1/2}L_K^{1/2} = L_K$, where $(L_K f)(x,y) = \int \mathbf{K}(x,y;x',y')f(x',y')$ is defined by a (continuous, symmetric) positive semi–definite $\mathbf{K}$ [94].

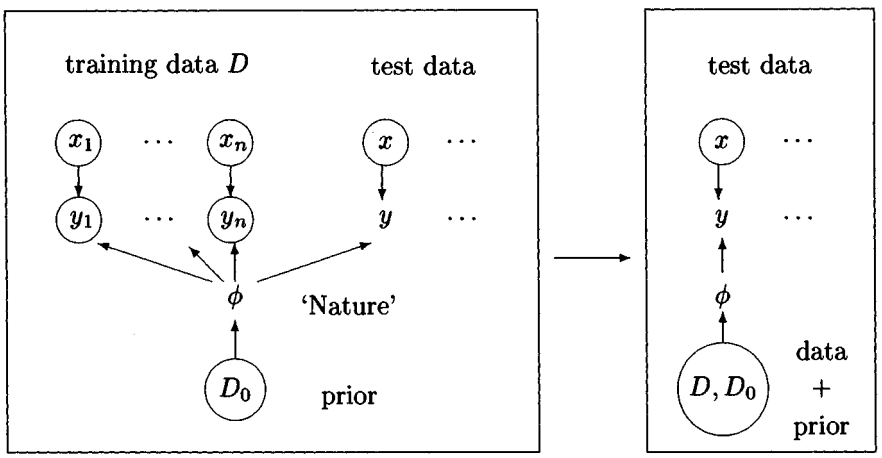


Figure 2.6: Graphical representation of learning. Given training data $D = (x_D, y_D)$ and *a priori* information D_0 we are interested in predicting outcome y for a new situation x. The graphical representation of the model $p(y, y_D, \phi | x, x_D, D_0) = p(y|x,\phi)\ p(\phi|D_0) \prod_{i=1}^{n} p(y_i|x_i,\phi)$ is shown on the left. Circles indicate measured variables, i.e., the available data D, D_0. Bayes' theorem (2.60) is then used to obtain the posterior density $p(\phi|D, D_0) \propto p(y_D|x_D,\phi)p(\phi|D_0)$ from which the predictive density (2.64) for test data (x, y) can be calculated by integrating (marginalizing) over ϕ, i.e., $p(y|x, f) = p(y|x, D, D_0) = \int d\phi\, p(y|x,\phi)\ p(\phi|D, D_0)$.

the *prior predictive probability density,*

$$p(y|x, D_0) = \int d\phi\, p(\phi|D_0)\, p(y|x,\phi) = < p(y|x,\phi) >_{\Phi|D_0}, \tag{2.63}$$

where $< \cdots >_{\Phi|D_0}$ denotes the expectation under the prior $p(\phi|D_0)$. After having received data, this is replaced by the *(posterior) predictive probability density*

$$p(y|x, f) = p(y|x, D, D_0) = \int d\phi\, p(\phi|f)\, p(y|x,\phi) = < p(y|x,\phi) >_{\Phi|f}, \tag{2.64}$$

$< \cdots >_{\Phi|f}$ denoting the expectation under the posterior $p(\phi|f) = p(\phi|D, D_0)$. The basic steps of Bayesian learning, calculation of the posterior and of the posterior predictive density are shown graphically in Fig. 2.6.

Note that $p(y|x, f)$ and $p(y|x, D_0)$ are in the convex hull spanned by the possible states of Nature $\phi \in \Phi$, and are not necessarily equal to one of the $p(y|x,\phi)$. (In case of general nonparametric density estimation all likelihoods are possible states of Nature.) The situation is illustrated in Fig. 2.7.

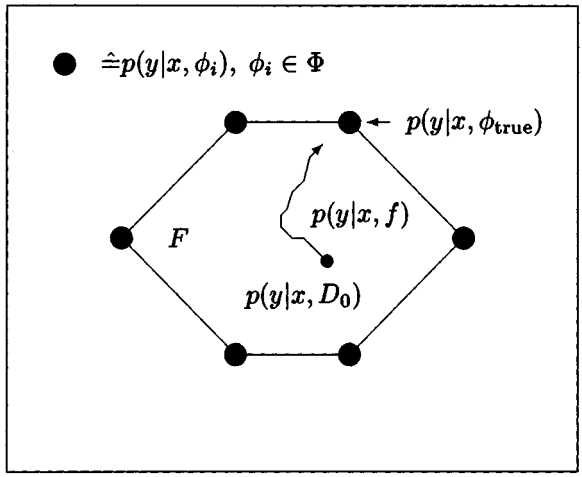


Figure 2.7: Evolution of the predictive density. The predictive density $p(y|x,f)$ for a state of knowledge $f = f(D, D_0, D_S)$ is in the convex hull spanned by the possible states of Nature ϕ_i characterized by the likelihoods $p(y|x,\phi_i)$. During learning, the actual predictive density $p(y|x,f)$ tends to move stochastically from the prior predictive density $p(y|x,D_0)$ towards that extremal point $p(y|x,\phi_{\text{true}})$ which represents the 'true' state of Nature.

During learning the predictive density $p(y|x,f)$ tends to approach the true $p(y|x,\phi)$. Because the training data are random variables, this approach is stochastic. There exists an extensive literature analyzing the stochastic properties of learning and generalization from a statistical mechanics perspective [143–145, 376, 511, 527].

When relating empirical learning to the framework of statistical field theory, it is interesting to note that using (2.60) the predictive density can be written as the ratio of two correlation functions under the prior $p(\phi)$,

$$p(y|x,f) = < p(y|x,\phi) >_{\Phi|f} \tag{2.65}$$

$$= \frac{< p(y|x,\phi) \prod_i p(y_i|x_i,\phi) >_{\Phi|D_0}}{< \prod_i p(y_i|x_i,\phi) >_{\Phi|D_0}} \tag{2.66}$$

$$= \frac{\int d\phi\, p(y|x,\phi)\, e^{-\beta E_{\text{comb}}}}{\int d\phi\, e^{-\beta E_{\text{comb}}}}. \tag{2.67}$$

Here $< \cdots >_{\Phi|D_0}$ denotes the expectation under the prior density $p(\phi) = p(\phi|D_0)$, and the *combined likelihood and prior energy* E_{comb} collects the ϕ–dependent energy and free energy terms

$$E_{\text{comb}} = \sum_i E(y_i|x_i,\phi) + E(\phi|D_0) - F(Y_D|x_D,\phi), \tag{2.68}$$

with

$$F(Y_D|x_D,\phi) = -\frac{1}{\beta}\ln Z(Y_D|x_D,\phi). \tag{2.69}$$

Going from Eq. (2.66) to Eq. (2.67) the normalization factor $Z(\Phi|D_0)$ appearing in numerator and denominator has been canceled.

2.1.5 Mutual information and learning

The aim of learning is to generalize the information obtained from training data to non–training situations. For such a *generalization* to be possible, there must exist a relation between the likelihoods $p(y_i|x_i,\phi)$ for training and for non–training data, provided by *a priori* information.

One possibility to quantify the relation between two random variables y_1 and y_2, representing for example training and non–training data, is to calculate their *mutual information*, defined as

$$M(Y_1,Y_2) = \sum_{y_1\in Y_1, y_2\in Y_2} p(y_1,y_2)\ln\frac{p(y_1,y_2)}{p(y_1)p(y_2)}. \tag{2.70}$$

It is also instructive to express the mutual information in terms of the entropy (2.35), which, for a probability function $p(y)$, is defined as

$$H(Y) = -\sum_{y\in Y} p(y)\ln p(y). \tag{2.71}$$

Hence

$$M(Y_1,Y_2) = H(Y_1) + H(Y_2) - H(Y_1,Y_2), \tag{2.72}$$

and $M = 0$ for independent y_i.

To have a compact notation for a family of predictive probabilities or densities $p(y_i|x_i,f)$ we choose a vector $x = (x_1,x_2,\cdots)$ consisting of all possible values x_i and corresponding vector of random variables $y = (y_1,y_2,\cdots)$, hence we can write

$$p(y|x,f) = p(y_1,y_2,\cdots|x_1,x_2,\cdots,f). \tag{2.73}$$

We now characterize a state of knowledge f corresponding to predictive density $p(y|x,f)$ by the mutual information of the components y_i of y. Thus we generalize the definition (2.70) from two random variables to a random vector y with components y_i, given vector x with components x_i, and obtain the conditional mutual information

$$M(Y|x,f) = \sum_{y_1\in Y_1, y_2\in Y_2,\cdots} p(y|x,f)\ln\frac{p(y|x,f)}{\prod_j p(y_j|x_j,f)} \tag{2.74}$$

or

$$M(Y|x,f) = \left(\sum_i H(Y_i|x_i,f) - H(Y|x,f)\right) \tag{2.75}$$

in terms of conditional entropies

$$H(Y_i|x_i,f) = -\sum_{y_i\in Y_i} p(y_i|x_i,f)\ln p(y_i|x_i,f), \tag{2.76}$$

$$H(Y|x,f) = -\sum_{y_1\in Y_1,y_2\in Y_2,\cdots} p(y|x,f)\ln p(y|x,f), \tag{2.77}$$

where for continuous y_i the sums have to be replaced by integrals. In case x is not given but sampled with density $p(x)$, it is useful to average the conditional mutual information and conditional entropy by including an integral $\int dx\, p(x)$ in the above formulae.

It is clear from Eq. (2.74) that probabilities or densities which factorize

$$p(y|x,f) = \prod_i p(y_i|x_i,f), \tag{2.78}$$

have mutual information zero. Hence such *factorial states* do not allow any generalization from training to non–training data. A special example of factorial states are the possible states of Nature (pure states) ϕ, which factorize according to the definition of our model

$$p(y|x,\phi) = \prod_i p(y_i|x_i,\phi). \tag{2.79}$$

Thus pure states do not allow any further generalization. This is consistent with the fact that pure states represent the natural end points of any learning process.

It is interesting to see that there are also other states f for which the predictive density factorizes. Indeed, from Eq. (2.79) together with Eq. (2.64) it follows that any (prior or posterior) $p(\phi)$ which factorizes in x leads to a factorial state,

$$p(\phi) = \prod_i p(\phi(x_i)) \Rightarrow p(y|x,f) = \prod_i p(y_i|x_i,f). \tag{2.80}$$

This means generalization or non–local learning between different x_i, is impossible when starting from a *factorial prior* as in (2.80) (see Fig. 2.8). A prior (2.80) which factorizes in x allows only local learning, i.e., learning with respect to the same x–value.[3] Introducing a prior on $\phi(x,y) = p(y|x,\phi)$ which factorizes exactly in x and y is in general not possible. This is due to the normalization constraint $\int dy\, p(y|x,\phi) = 1$ which is nonlocal in y. But if many y–values contribute with similar weights to the normalization this constraint is weak and a prior factorial in x and y can be a good approximation. Note also

[3]One can however always introduce a more complete set of x–variables, e.g., by introducing a unique label (time, location) for each measurement, so that training data D necessarily contain only different values x_i.

that a state of knowledge $p(\phi|f) = \delta(\phi' - \phi) = \prod_{x_i,y_j} \delta(\phi'(x_i, y_j) - \phi(x_i, y_j))$ setting ϕ to a pure state ϕ' is always factorial in x and y.

A factorial prior provides a very clear reference for analyzing the role of *a priori* information in learning. In particular, with respect to a prior, factorial in the variables x_i, learning may be decomposed into two steps, one increasing, the other lowering mutual information:

1. Starting from a factorial prior, new *non–local data* D_0 (typically called *a priori* information) produce a new non–factorial state with non–zero mutual information.

2. *Local data* D (typically called training data) reduce stochastically the mutual information. Hence learning with local data corresponds to a *stochastic decay of mutual information.*

Pure states, i.e., the extremal points in the space of possible predictive densities, are in general probabilistic with finite variance and can be further decomposed into components ψ with lower variance,

$$p(y_i|x_i, \phi) = \int d\psi \, p(y_i|x_i, \psi) \, p(\psi|\phi). \tag{2.81}$$

'Freezing' one of the components ψ, for example, by using improved measurement devices, leaves a new pure state ψ with lower variance. Then, imposing a non–factorial prior $p(\psi|\phi)$ on the new, finer hypotheses ψ enables again non–local learning with local data, leading asymptotically to one of the new pure states $p(y_i|x_i, \psi)$.

Let us exemplify the stochastic decay of mutual information by a simple numerical example. Because the mutual information requires the integration over all y_i variables, we choose a problem with only two of them, y_a and y_b, corresponding to two values x_a and x_b of x. We thus consider a model with four states of Nature ϕ_l, $1 \leq l \leq 4$, with Gaussian likelihood $p(y|x, \phi) = (\sqrt{2\pi}\sigma)^{-1} \exp\left(-(y - \phi_i(x))^2/(2\sigma^2)\right)$ and local mean values $\phi_l(x_j) = \pm 1$. A hypothesis ϕ_l is then fully characterized by its two local mean values $\phi_l(x_a) = \phi_{a,l}$, $\phi_l(x_b) = \phi_{b,l}$ for y_a and y_b. Selecting one of the four possible states as 'true' state of Nature ϕ, we sample 50 data points $D_i = (x_i, y_i)$, $i = 1, \ldots, 50$, from the corresponding Gaussian likelihood using $p(x_a) = p(x_b) = 0.5$. Then, starting from a given, factorial or non–factorial, prior $p(\phi|D_0)$ we sequentially update the predictive density (2.64),

$$p(y|x, D_i, \cdots, D_0) = \sum_{l=1}^{4} p(y|x, \phi_l) \, p(\phi_l|D_i, \cdots, D_0), \tag{2.82}$$

by calculating the posterior according to (2.60)

$$p(\phi_l|D_i, \cdots, D_0) = \frac{p(y_i|x_i, \phi_l) \, p(\phi_l|D_{i-1}, \cdots, D_0)}{p(y_i|x_i, D_{i-1}, \cdots, D_0)}. \tag{2.83}$$

A factorial prior takes for this example the form $p(\phi_l|D_0) = p(\phi_{a,l}|D_{a,0})$ $\times p(\phi_{b,l}|D_{b,0})$ with $D_{a,0}$, $D_{b,0}$ denoting the *a priori* information under x_a or x_b, respectively. Decomposing the sum in Eq. (2.82) for the two components of ϕ according to $\sum_{l=1}^{4} = \sum_{l_a=1}^{2}\sum_{l_b=1}^{2}$ it is easily seen that factorial states remain factorial under local data

$$\begin{aligned} p(y|x,D,D_0) &= \sum_{l=1}^{4} p(y_a,y_b|x_a,x_b,\phi_l)\frac{p(\phi_l|D_0)}{p(y_D|x_D,D_0)}p(y_D|x_D,\phi_l) \\ &= \sum_{l_a=1}^{2} p(y_a|x_a,\phi_{a,l_a})\frac{p(\phi_{a,l_a}|D_{a,0})}{p(y_{D_a}|x_{D_a},D_{a,0})}p(y_{D_a}|x_{D_a},\phi_{a,l_a}) \\ &\quad \times \sum_{l_b=1}^{2} p(y_b|x_b,\phi_{b,l_b})\frac{p(\phi_{b,l_b}|D_{b,0})}{p(y_{D_b}|x_{D_b},D_{b,0})}p(y_{D_b}|x_{D_b},\phi_{b,l_b}) \\ &= p(y_a|x_a,D_a,D_{a,0})\,p(y_b|x_b,D_b,D_{b,0}), \end{aligned} \tag{2.84}$$

where $D_a = (x_{D_a}, y_{D_a})$ and $D_b = (x_{D_b}, y_{D_b})$ denote the data for $x_i = x_a$ or $x_i = x_b$, respectively. Fig. 2.9 shows that indeed the mutual information decays rapidly. Depending on the training data, a wrong hypothesis ϕ_l may still survive the decay of mutual information. Having arrived at a factorial state, further learning can only be local. Data points for x_i can then only influence the predictive density for the corresponding y_i and do not allow generalization to the other y_j with $j \neq i$. For a factorial prior $p(\phi_l) = p(\phi_l(x_a))p(\phi_l(x_b))$ learning is thus local from the very beginning. Only very small numerical random fluctuations of the mutual information occur and are quickly eliminated by learning. Thus the predictive density moves through a sequence of factorial states.

2.1.6 Maximum A Posteriori Approximation (MAP)

In most applications the high dimensional ϕ–integral over model states in Eq. (2.64) or Eq. (2.67) can not be performed exactly. The two most common methods to calculate the ϕ–integral approximately are *Monte Carlo integration* [33,48,150,152,153,195,204,278,337,363,364,370,414,415,426,460,477, 481,524] and *saddle point approximation* [33,35,59,101,150,162,291,355,365, 424,477,541]. The latter approach will be studied in the following.

For that purpose, we expand E_{comb} of Eq. (2.68) with respect to ϕ around some ϕ^*

$$\begin{aligned} E_{\text{comb}}(\phi) &= e^{(\Delta\phi|\nabla)}E(\phi)\Big|_{\phi=\phi^*} \\ &= E_{\text{comb}}(\phi^*) + \langle\Delta\phi|\,\nabla E_{\text{comb}}(\phi^*)\rangle + \frac{1}{2}\langle\Delta\phi|\mathbf{H}(\phi^*)|\,\Delta\phi\rangle + \cdots \end{aligned} \tag{2.85}$$

with $\Delta\phi = (\phi - \phi^*)$, gradient ∇ (not acting on $\Delta\phi$), Hessian $\mathbf{H}$, and $(\cdots|\cdots)$, $\langle\cdots|\cdots|\cdots\rangle$ denoting scalar products or matrix elements of real–symmetric

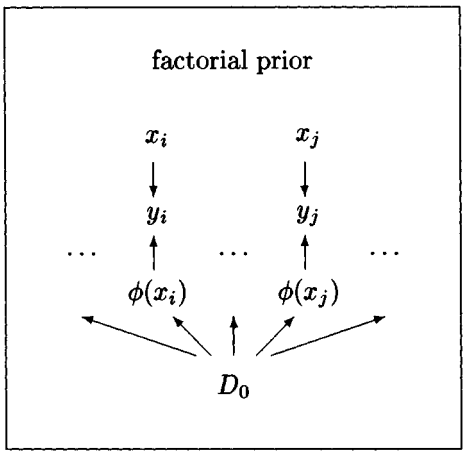


Figure 2.8: Graphical representation of a probabilistic model with factorial prior over x, i.e., $p(\phi|D_0) = \prod_k' p(\phi(x_k)|D_0)$, where the product $\prod_k'$ includes all values x_k only once. For example, in regression with Gaussian $p(y|x,\phi)$ and fixed variance, $\phi(x)$ represents the regression function, i.e., the mean of y at x. The corresponding model $p(y,\phi|x,D_0) = \left(\prod_k' p(\phi(x_k)|D_0)\right) \prod_i p(y_i|x_i,\phi_i)$ factorizes completely in x. A factorial prior represents a natural 'no-knowledge' prior: Under a factorial prior data on x_i contain no information for a different $x_j \neq x_i$, hence nonlocal learning or generalization to new x-values is not possible.

(hermitian) operators, respectively. Thus for example,

$$\langle \Delta\phi |\, \nabla E_{\text{comb}}(\phi^*) \rangle = \int dx\, dy\; (\phi(x,y) - \phi^*(x,y))\, \nabla E_{\text{comb}}(\phi^*)(x,y). \quad (2.86)$$

For functions $\phi(x,y)$ this yields

$$\nabla E_{\text{comb}}(x,y;\phi^*) = \left.\frac{\delta E_{\text{comb}}(\phi)}{\delta\phi(x,y)}\right|_{\phi=\phi^*}, \quad (2.87)$$

$$\mathbf{H}(x,y,x',y';\phi^*) = \left.\frac{\delta^2 E_{\text{comb}}(\phi)}{\delta\phi(x,y)\delta\phi(x',y')}\right|_{\phi=\phi^*}, \quad (2.88)$$

where the derivatives are functional (Fréchet) derivatives [58, 80, 210, 240] or partial derivatives for discrete x, y.

Choosing ϕ^* to be the location of a local minimum of $E_{\text{comp}}(\phi)$ the linear term in (2.85) vanishes. The second order term of the log-posterior in Eq. (2.67) and Eq. (2.64) includes the Hessian and corresponds to a Gaussian

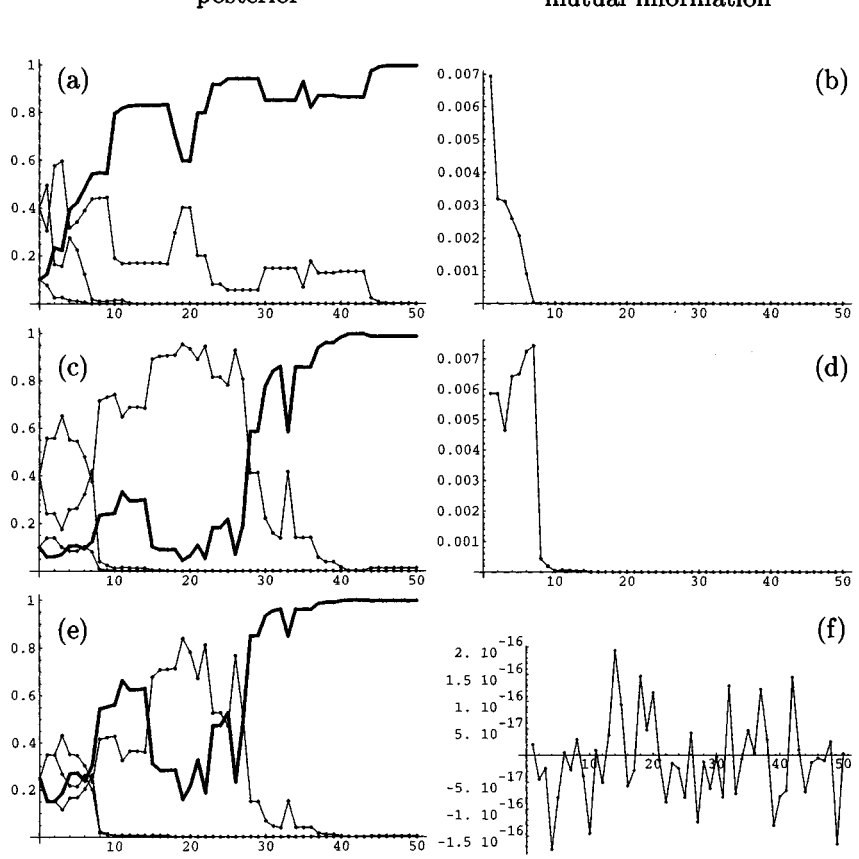


Figure 2.9: The decay of mutual information during learning. Model with 4 possible states ϕ_l representing Gaussian likelihoods $p(y_i|x_i, \phi_l)$ with means ± 1 for two different x_i values. Shown are posterior probabilities $p(\phi_l|f)$ (a, c, e, on the left hand side, the posterior of the true ϕ_l is shown by a thick line) and mutual information $M(y)$ (b, d, f, on the right) during learning from 50 training data. (a, b): The mutual information decays during learning and quickly becomes practically zero. (c, d): For 'unlucky' training data a wrong hypothesis ϕ_i can dominate at the beginning. Nevertheless the mutual information decays and the correct hypothesis is finally found through 'local' learning, where data for x_i do not contain information on data for the other $x_j \neq x_i$. (e, f): Starting with a factorial prior the mutual information is and remains zero, apart from artificial numerical fluctuations. For (e, f) the same random data have been used as for (c, d).

integral over ϕ which can be solved analytically

$$\int d\phi\, e^{-\frac{1}{2}(\Delta\phi, \mathbf{H}\Delta\phi)} = (2\pi)^{\frac{d}{2}} (\det \mathbf{H})^{-\frac{1}{2}}, \tag{2.89}$$

for a d–dimensional ϕ–integral. Using the same approximation for the ϕ–integral in the numerator of Eq. (2.67), expanding also $p(y|x,\phi)$ around ϕ^* and restricting to the first (ϕ–independent) term $p(y|x,\phi^*)$ of that expansion, the factor (2.89) cancels, even for infinite d. The result is the zero order term of an expansion of the predictive density in powers of $1/\beta$. Higher order contributions can be calculated by using Wick's theorem [59,101,220,221,251, 291,351,365,500,541]. The final approximate result for the predictive density (2.64) is very simple and intuitive

$$p(y|x,f) \approx p(y|x,\phi^*), \tag{2.90}$$

with

$$\phi^* = \mathrm{argmin}_{\phi\in\Phi} E_{comb} = \mathrm{argmax}_{\phi\in\Phi}\, p(\phi|f) = \mathrm{argmax}_{\phi\in\Phi}\, p(y_D|x_D,\phi) p(\phi|D_0). \tag{2.91}$$

This saddle point (or Laplace) approximation is called *Maximum A Posteriori Approximation* (MAP). The relation between a full Bayesian approach and MAP is strictly analogous to the relation between quantum (field) theory (in path integral formulation) and classical physics.

It is also useful to note that the fact that the inverse Hessian is the covariance for ϕ of the Gaussian approximation (2.89) allows to read off the *error bars* for an approximation ϕ^*. Indeed, we have

$$\mathrm{cov}(\phi^*(x,y), \phi^*(x',y')) = \mathbf{H}^{-1}(x,y;x',y'), \tag{2.92}$$

and thus obtain for the standard deviation (error bar) at $\phi^*(x,y)$

$$\sigma(\phi^*(x,y)) = \sqrt{\mathbf{H}^{-1}(x,y;x,y)}. \tag{2.93}$$

Notice that the MAP solution ϕ^* also maximizes the integrand of the evidence of the data y_D, being proportional to $p(\phi|D,D_0)$,

$$p(y_D|x_D,D_0) = \int d\phi\, p(y_D|x_D,\phi) p(\phi|D_0). \tag{2.94}$$

This fact results from the assumption that $p(y|x,\phi)$ in Eq. (2.64) is slowly varying at the stationary point and does not influence the location ϕ^* of the saddle point in a Maximum A Posteriori Approximation for the predictive density. For (functional) differentiable E_{comb}, Eq. (2.91) yields the stationarity equation,

$$\frac{\delta E_{\mathrm{comb}}(\phi)}{\delta\phi(x,y)} = 0. \tag{2.95}$$

The functional E_{comb} which includes training and prior data terms is also known as (regularized) *error functional* for ϕ. (The prior terms correspond to the regularization or stabilizer terms in a Frequentist approach, see Section 2.2.5.)

In practice a saddle point approximation is useful if the posterior is well approximated by a Gaussian centered at the maximum. Asymptotical convergence results for MAP–approaches in the limit $\beta \to \infty$ (with the same β for the prior and data term) can be obtained, if E_{comb} approaches a constant in this limit, so that $e^{-\beta E_{\text{comb}}}$ becomes sharply peaked. (See [90, 531].) For the likelihood term this means that

$$-\frac{1}{\beta}\sum_i L(y_i|x_i,\phi) \tag{2.96}$$

has to become β–independent for $\beta \to \infty$. If, for example, $\frac{1}{n}\sum_i L(y_i|x_i,\phi) = K_c$ (which corresponds to the conditional Kerridge inaccuray of Eq. (2.33)) converges for a large number n of training data and if the prior energy scales with n so that $E(\phi|D_0)$ converges towards a constant (e.g., vanishes) for large n, then, identifying β with n, the low temperature limit $1/\beta \to 0$ can be interpreted as large data limit $n \to \infty$,

$$\beta E_{\text{comb}} \to nE_{\text{comb}} = n\left(-\frac{1}{n}\sum_i L(y_i|x_i,\phi) + E(\phi|D_0)\right). \tag{2.97}$$

For regression problems with a Gaussian likelihood $p(y|x,\phi)$ one can similarly introduce a temperature $1/\beta = \sigma^2$, corresponding to the variance,

$$\beta E_{\text{comb}} \to \frac{1}{\sigma^2}E_{\text{comb}} = \frac{1}{\sigma^2}\left(\frac{1}{2}\sum_i |y_i - \phi(x_i)|^2 + E(\phi|D_0)\right). \tag{2.98}$$

Thus for a Gaussian prior, asymptotic convergence results for $\sigma^{-2} \to \infty$ can be obtained if the prior variance is scaled simultaneously with the data variance, so that $E(\phi|D_0)$ becomes asymptotically σ^2–independent.

We should also remark that for continuous x, y the stationary solution ϕ^* needs not to be a typical representative of the process $p(\phi|f)$. A common example is a Gaussian stochastic process $p(\phi|f)$ with prior energy $E(\phi|D_0)$ related to some smoothness measure of ϕ expressed by derivatives of $p(y|x,\phi)$. Then, even if the stationary ϕ^* is smooth as a function of x, y, this needs not to be the case for a typical ϕ sampled according to $p(\phi|f)$. For Brownian motion, for instance, a typical sample path is not even differentiable (but continuous) while the stationary path, i.e., the mean, is smooth. Thus, for continuous variables only expressions like $\int d\phi\, e^{-\beta E(\phi)}$ can be given an exact meaning as a Gaussian measure, defined by a given covariance with existing normalization factor, but not the separate expressions $d\phi$ and $E(\phi)$ as such [111, 142, 183, 309, 491, 507].

2.1.7 Normalization, non–negativity, and specific priors

Density estimation problems are characterized by their normalization and non–negativity constraints for $p(y|x,\phi)$. The prior density $p(\phi|D_0)$ can thus only be non–zero for such ϕ for which $p(y|x,\phi)$ is non–negative and normalized over y for all x.[4] While the non–negativity constraint is local with respect to x and y, the normalization constraint is nonlocal with respect to y. Thus implementing a normalization constraint leads to nonlocal and in general non–Gaussian priors.

For classification problems having discrete y–values (classes), the normalization constraint requires only to sum over the different classes, which often can be done, and a Gaussian prior structure with respect to the x–dependency is not altered [521]. For general density estimation problems with continuous y, however, the loss of the Gaussian structure with respect to y is more severe, because non–Gaussian functional integrals can in general not be performed analytically. On the other hand, solving the learning problem numerically by discretizing the y– and x–variables, the normalization term is typically not a severe complication.

The prior $p(\phi|D_0)$ has thus to implement the non–negativity and normalization conditions

$$Z_X(x,\phi) = \int dy\, p(y|x,\phi) = 1, \qquad \forall x \in X, \forall \phi \in \Phi \tag{2.99}$$

$$p(y|x,\phi) \geq 0, \qquad \forall y \in Y, \forall x \in X, \forall \phi \in \Phi. \tag{2.100}$$

It is useful to isolate the normalization and non–negativity constraint from the problem–specific priors. Introducing the *specific prior information* $\tilde{D}_0$ so that $D_0 = \{\tilde{D}_0, \text{normalized, non-negative}\} = \{\tilde{D}_0, \text{norm, nn}\}$, we have with (2.58)

$$p(\phi|\tilde{D}_0, \text{norm}, \text{nn}) = \frac{p(\text{norm}, \text{nn}|\phi) p(\phi|\tilde{D}_0)}{p(\text{norm}, \text{nn}|\tilde{D}_0)}, \tag{2.101}$$

with deterministic, $\tilde{D}_0$–independent

$$p(\text{norm}, \text{nn}|\phi, \tilde{D}_0) = p(\text{norm}, \text{nn}|\phi) \tag{2.102}$$

$$= p(\text{norm}|\phi) p(\text{nn}|\phi) = \delta(Z_X - 1) \prod_{xy} \Theta\Big(p(y|x,\phi)\Big), \tag{2.103}$$

and step function Θ. (The density $p(\text{norm}|\phi)$ is normalized over all possible values of Z_X, and $p(\text{nn}|\phi)$ over all possible sign combinations.) The denominator $p(\text{norm}, \text{nn}|\tilde{D}_0)$ is ϕ–independent. We call

$$p(\phi|\tilde{D}_0) = e^{-E(\phi|\tilde{D}_0) + F(\Phi|\tilde{D}_0)} \tag{2.104}$$

[4]Similarly, when solving for a distribution function, i.e., the integral of a density, the non–negativity constraint is replaced by monotonicity and the normalization constraint by requiring the distribution function to be 1 on the upper boundary (see Section 3.5.3).

the *specific prior* and $E(\phi|\tilde{D}_0)$ the specific prior energy, choosing from now on $\beta = 1$. The free energy $F(\Phi|\tilde{D}_0)$ does not depend on ϕ. Written in terms of the specific prior, the posterior reads

$$\begin{aligned} p(\phi|f) &= C\, p(\phi|\tilde{D}_0) \left(\prod_i p(y_i|x_i,\phi) \right) \\ &\times \left(\prod_{x'} \delta \left(\int dy\, p(y|x',\phi) - 1 \right) \right) \\ &\times \left(\prod_{x''y''} \Theta\Big(p(y''|x'',\phi) \Big) \right), \end{aligned} \tag{2.105}$$

with ϕ–independent $C = p(\text{norm}, \text{nn}|\tilde{D}_0)^{-1}$. As the normalization and non–negativity constraints are implemented explicitly in Eq. (2.105) we are now free to choose for the specific prior a convenient form. Typically, in nonparametric models this is a Gaussian process prior

$$\begin{aligned} p(\phi|\tilde{D}_0) &= \frac{e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t\rangle}}{\int d\phi'\, e^{-\frac{1}{2}\langle \phi' - t|\mathbf{K}|\phi' - t\rangle}} \\ &= \left(\det \frac{\mathbf{K}}{2\pi} \right)^{\frac{1}{2}} e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t\rangle} \propto e^{-E(\phi|\tilde{D}_0)}, \end{aligned} \tag{2.106}$$

with specific prior information $\tilde{D}_0 = (t, \mathbf{K})$, representing the mean function $t(x,y)$ and inverse covariance $\mathbf{K}$, and

$$\langle \phi - t\,|\mathbf{K}|\,\phi - t\rangle =$$

$$\int dx\, dy\, dx'\, dy' (\phi(x,y) - t(x,y)) \mathbf{K}(x,y;x',y') (\phi(x',y') - t(x',y')). \tag{2.107}$$

Gaussian process priors as in Eq. (2.106) are in wide–spread use for example in nonparametric regression [322, 520, 522], the Bayesian approach to classical inverse problems [222, 478], classification [520, 521], general density estimation [296, 299, 456]; they will also be used in Chapter 7 for Bayesian inverse quantum theory. A Bayesian MAP approach is often technically equivalent to a classical regularization approach [507]. In that case Gaussian priors correspond to the choice of quadratic regularizers in regularization approaches [222, 262, 482], see Section 2.2.5. The interpretation of a regularization term and a prior term can, however, be different, one representing an additional target or cost function, the other additional information on Nature (cf. Section 2.2.6 and [181, 291]).

Typical choices for inverse prior covariances $\mathbf{K}$ are differential operators which enforce the function ϕ to be smooth. For example, taking for a one–dimensional regression problem as $\mathbf{K}$ the negative Laplacian multiplied by

some factor λ and choosing a zero mean, $t = 0$, yields a prior energy $E(\phi|\tilde{D}_0) = E(\phi|D_0)$ for the regression function $\phi(x)$

$$\frac{1}{2}\langle \phi - t\,|\mathbf{K}|\,\phi - t\rangle = -\frac{\lambda}{2}\int dx\,\phi(x)\,\frac{\partial^2}{\partial x^2}\phi(x) = \frac{\lambda}{2}\int dx\,\left(\frac{\partial\phi(x)}{\partial x}\right)^2, \quad (2.108)$$

assuming vanishing boundary terms. In statistical applications one often prefers inverse covariance operators with higher derivatives to obtain smoother regression functions [8,162,183,263,453,507,519]. An example of such a prior with higher derivatives is a 'Radial Basis Function' prior with the pseudo–differential operator

$$\mathbf{K} = \lambda \exp\left(-\sigma_0^2\Delta/2\right) \quad (2.109)$$

as inverse covariance [124,159,399,400,507].

A Gaussian as in (2.106) maximizes *entropy* under the constraints

$$\begin{aligned} 1 &= \int d\phi\, p(\phi), \\ t(x,y) &= \int d\phi\, \phi(x,y) p(\phi), \\ \mathbf{K}^{-1}(x,y;x',y') &= \int d\phi\, (\phi(x,y) - t(x,y))(\phi(x',y') - t(x',y'))p(\phi). \end{aligned} \quad (2.110)$$

Technically this maximization with constraints is done by introducing a Lagrange parameter α for the normalization constraint, a Lagrange parameter function $\beta(x,y)$ for the mean function, and Lagrange parameter matrix or operator $\gamma(x,y;x',y')$ for the covariance [458]. As mentioned in Section 2.1.2, the prior (2.106), like every probability density, also represents a maximum entropy solution $p \propto e^{-\beta E(\phi|\tilde{D}_0)}$ ((Boltzmann–)Gibbs distribution, generalized canonical ensemble) for fixed expectation of $E(\phi|\tilde{D}_0)$ (fixed in this case to the value which corresponds to $\beta = 1$). Analogously, maximizing the Kullback–Leibler distance or relative entropy (2.47) with reference density $p(\phi|norm, nn)$ leads to $p(\phi|D_0)$.

In particular, if prior information on the mean function and covariance of $p(\phi|\tilde{D}_0)$ is not available, a standard choice is a pure maximum entropy prior [187,458] (cf. Section 2.1.2 and Eq. (2.47)),

$$p(\phi) \propto e^{\alpha S(\phi)} \quad (2.111)$$

where $\alpha > 0$ and $S(\phi)$ stands for (a slight generalization of) the Shannon–Jaynes entropy

$$S(\phi) = \int dy\,dx\,\left(p(y|x,\phi) - m(x,y) - p(y|x,\phi)\ln\frac{p(y|x,\phi)}{m(x,y)}\right). \quad (2.112)$$

For normalized likelihood and normalized m the first two terms cancel. The function $m(x,y)$ represents a reference function or integration measure. In the

absence of further data, the prior is maximized for $\phi(x,y) = m(x,y)$. Thus the role of $m(x,y)$ as a reference function is similar to the role of the mean function $t(x,y)$ in the Gaussian prior (2.106). A pure maximum entropy prior, however, is not well suited to implement nonlocal conditions like smoothness, if not combined with nonlocal data terms. Hence a maximum entropy prior is mainly useful if formulated for an underlying function from which the data are obtained by nonlocal noise or averaging processes, like in the classical inverse problems of the kind of Eq. (2.12). A maximum entropy prior may, for example, be formulated for an original image while the observed noisy data are obtained by averaging pixels of the original image over some neighborhood [187,458]. Special examples are some ensembles of random matrices which are obtained as maximum entropy solutions, fixing the expectation of the trace of the squared matrix [334].

A specific prior of the form $p(\phi|\tilde{D}_0)$ as in Eq. (2.104) is implemented by controlling the parameters $\tilde{D}_0$ of a ϕ–generating process and represents a *controlled prior*. Alternatively, a specific prior in likelihood form $p(\tilde{D}_0|\phi)$ can be interpreted as a generalized measurement of ϕ with result $\tilde{D}_0$ [291]. This corresponds to a *measured prior* which is related to the full prior $p(\phi|D_0)$ according to

$$\begin{aligned} p(\phi|D_0) &= p(\phi|\tilde{D}_0, \mathrm{norm}, \mathrm{nn}) \\ &= \frac{p(\tilde{D}_0|\phi)\, p(\phi|\mathrm{norm}, \mathrm{nn})}{p(\tilde{D}_0|\mathrm{norm}, \mathrm{nn})} \\ &= \frac{p(\tilde{D}_0|\phi)\, p(\mathrm{norm}, \mathrm{nn}|\phi)\, p(\phi)}{p(\tilde{D}_0|\mathrm{norm}, \mathrm{nn}) p(\mathrm{norm}, \mathrm{nn})}. \end{aligned} \tag{2.113}$$

Here $p(\phi)$ is a higher level prior which can be chosen uniformly if all *a priori* information besides normalization and non–negativity is represented by $\tilde{D}_0$. Since the (generalized) likelihood $p(\tilde{D}_0|\phi) = e^{-E(\tilde{D}_0|\phi)}/Z$ is conditional on ϕ, the normalization factor $Z = \int d\tilde{D}_0\, e^{-E(\tilde{D}_0|\phi)}$ remains in general ϕ–dependent and must be included when maximizing the posterior with respect to ϕ (see Section 2.1.6). It is a special property of Gaussian specific priors with fixed covariances that measured and controlled interpretation coincide. Let us interpret Gaussian specific prior data $\tilde{D}_0 = (t, \mathbf{K})$ as having measured a function $t(x,y)$ under Gaussian noise with covariance $\mathbf{K}^{-1}$. Then $p(\tilde{D}_0|\phi) = p(t|\mathbf{K}, \phi)\, p(\mathbf{K})$, and one finds for fixed $\mathbf{K}$, due to the fact that the normalization of a Gaussian is independent of the mean,

$$p(t|\phi, \mathbf{K}) = \frac{e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t\rangle}}{\int dt'\, e^{-\frac{1}{2}\langle \phi - t'|\mathbf{K}|\phi - t'\rangle}} = \left(\det \frac{\mathbf{K}}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t\rangle}, \tag{2.114}$$

which is equal to the controlled prior $p(\phi|\tilde{D}_0)$ of Eq. (2.106). Eq. (2.114) is a generalization of standard training data for Gaussian regression (where ϕ

stands for a regression function $\phi(x)$, see Sections 2.3, 3.7, and 3.8),

$$p(y|x,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\phi(x))^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\langle y-\phi|\mathbf{K}_x|y-\phi\rangle} \tag{2.115}$$

with $\mathbf{K}_x(x,x') = \sigma^{-2}\delta(x-x')$ and $y(x) = y$. Thus for Gaussian prior $p(t|\phi,\mathbf{K})$ with a ϕ–independent normalization, the specific prior energy in likelihood form becomes, analogous to Eq. (2.104),

$$p(t|\phi,\mathbf{K}) \propto e^{-E(t|\phi,\mathbf{K})}, \tag{2.116}$$

and specific prior energies can be interpreted both ways.

To maximize the posterior (2.105) the non–negativity and normalization conditions can be implemented by Lagrange multipliers. First, we note that the non–negativity constraint contributes to the gradient only at locations where $p(y|x,\phi) = 0$. In all other cases the gradient has no component pointing into forbidden regions and the non–negativity constraint is not active [40,134, 223,402]. Hence, if strict positivity is checked for the final solution, then it is not necessary to include extra non–negativity terms in the prior (see Section 3.2.1). For instance, a smooth interpolation of the likelihood at the data points, where $p(y_i|x_i,\phi)$ has to be positive by definition, typically leads to a strictly positive $p(y|x,\phi) > 0$, at least as long as no 'overshooting' occurs. For the sake of simplicity the non–negativity term is thus not included in the following.

Written in terms of the log–likelihood $L(x,y) = \ln p(y|x,\phi)$ the posterior (2.105) becomes

$$p(\phi|f) \propto e^{\sum_i L(y_i|x_i,\phi) - E(\phi|\tilde{D}_0)} \prod_{x\in X} \delta\left(\int dy\, e^{L(y|x,\phi)} - 1\right). \tag{2.117}$$

Writing the δ–functional in its Fourier representation

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, e^{ikx} = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} dk\, e^{-kx}, \tag{2.118}$$

so that

$$\delta\Big(\int dy\, e^{L(y|x,\phi)} - 1\Big) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\Lambda_X(x)\, e^{\Lambda_X(x)\left(1-\int dy\, e^{L(y|x,\phi)}\right)}, \tag{2.119}$$

and performing a saddle point approximation with respect to $\Lambda_X(x)$ (which is exact in this case) yields

$$p(\phi|f) \propto e^{\sum_i L(y_i|x_i,\phi) - E(\phi|\tilde{D}_0) + \int dx\, \Lambda^*_X(x)\left(1-\int dy e^{L(y|x,\phi)}\right)}. \tag{2.120}$$

Here the stationary $\Lambda^*_X(x)$ is the Lagrange multiplier vector (or function) to be determined by the normalization conditions for $p(y|x,\phi) = e^{L(y|x,\phi)}$. This

is equivalent to minimize the energy (2.68) with Lagrange multiplier function Λ_X

$$E_{\text{comb}}(\phi) = -\sum_i L(y_i|x_i,\phi) + E(\phi|\tilde{D}_0) + \int dx\, \Lambda_X(x) \left(\int dy e^{L(y|x,\phi)} - 1 \right). \tag{2.121}$$

In case the non–negativity constraint has to be included this can be done, for example, by writing the step functions in $p(\text{nn}|\phi) \propto \prod_{x,y} \Theta(p(y|x,\phi))$ as

$$\Theta(x-a) = \int_a^\infty d\xi \int_{-\infty}^\infty \frac{d\eta}{2\pi} e^{i\eta(\xi - x)}, \tag{2.122}$$

and solving the ξ–integral in saddle point approximation, or alternatively, by using Lagrange multipliers. Sometimes it can be useful for numerics to add besides the above Lagrange multiplier term some additional terms to the log–posterior which vanish for normalized $p(y|x,\phi)$.

2.1.8 Numerical case study: A fair coin?

To go through the typical Bayesian calculations we will choose one of the simplest possible examples: the coin [457]. For the coin the hidden variable we are interested in is the single number $h = p(\text{'head'})$, $0 \leq h \leq 1$, representing the probability (likelihood) to find 'head' if tossing the coin. In that case, the hidden variables ϕ are not a complete function $\phi(x,y)$ but just a normalized two–dimensional vector

$$\phi = \begin{pmatrix} \phi_0 \\ \phi_1 \end{pmatrix} = \begin{pmatrix} p(\text{'tail'}) \\ p(\text{'head'}) \end{pmatrix} = \begin{pmatrix} 1-h \\ h \end{pmatrix}, \tag{2.123}$$

with the number h as single degree of freedom. For a single trial it is not necessary here to include a visible independent variable x. (Equivalently we can say x is a constant.) Thus, the likelihood function $p(y|x,\phi)$ for tossing the coin once reads $p(y|x,\phi) = p(y|\phi) = p(y|h)$ with $p(y=1|h) = p(\text{'head'}) = h = \phi_1$ and $p(y=0|h) = p(\text{'tail'}) = 1-h = 1-\phi_1 = \phi_0$.

Finally, the visible, dependent variable y of a single trial is binary, as it only takes the values 1 or 0, representing 'head' or 'tail'. Assuming a constant probability $p(\text{'head'})$ over time, the ordering of independent y_i in the data $D = \{(y_i)|1 \leq i \leq n\}$ is irrelevant. Therefore we can condense the data D to the number k of observed 'heads' in n trials. With k playing the role of the observed dependent data y_D, we obtain the binomial likelihood

$$p(y_D|x_D,\phi) = p(D|h) = p(k|n,h) = \binom{n}{k} h^k (1-h)^{(n-k)}. \tag{2.124}$$

Writing $p(y_D|x_D,\phi) = p(D|h)$ we have treated $x_D = n$ as fixed. Clearly, one could also allow a varying n.

In this notation we have for the hidden variable $\phi_1 = h$

$$\begin{aligned} \phi_1 &= h \\ &= p(\text{'head'}) = p(k = 1|n = 1, h) \\ &= p(y = 1|h) = p(y = 1|x = \text{const.}, \phi). \end{aligned} \tag{2.125}$$

Now our aim is to calculate an estimate for h, given a number k for some n.

Having defined the likelihood model for the coin, it remains to select a prior model $p(h)$ to be able to enter the Bayesian formalism. Clearly, which prior is appropriate depends on the specific situation. The following three examples are priors representing three quite different circumstances.

1. Knowing 'nothing' about h (which includes not knowing that the data are generated by a coin or not knowing the object 'coin') it seems reasonable to start with a uniform prior

$$p(h) = \Theta(h)\Theta(1 - h), \tag{2.126}$$

where

$$\Theta(x) = \begin{cases} 1 \text{ for } x \geq 0 \\ 0 \text{ for } x < 0 \end{cases}. \tag{2.127}$$

2. However, knowing that we are dealing with a coin and being 'experts' in the tossing of coins from our everyday life, we probably will prefer priors centered around $h = 0{,}5$, for example, the following (nearly) Gaussian 'expert' prior

$$p(h) \propto \Theta(h)\Theta(1 - h)e^{-\frac{|h-0.5|^2}{2\sigma^2}}. \tag{2.128}$$

3. On the other hand, if gambling in Las Vegas, we may not trust a coin too much and prefer a 'Las Vegas' prior like

$$p(h) \propto \Theta(h)\Theta(1 - h)\left(e^{-\frac{|h-0.5|^2}{2\sigma^2}} + 2e^{-\frac{|h-1|^2}{2\sigma^2}}\right). \tag{2.129}$$

Having specified likelihood and prior model it is straightforward to calculate the posterior

$$p(h|D) = p(h|n, k) = \frac{p(k|n, h)\, p(h)}{p(k|n)}, \tag{2.130}$$

with norming constant

$$p(k|n) = \int_0^1 dh\, p(k|n, h)\, p(h). \tag{2.131}$$

The predictive probability for 'head' becomes the mean of the posterior

$$\begin{aligned} p_{\text{pred}}(\text{'head'}) &= p(\text{'head'}|D) \\ &= p(k'=1|n'=1,n,k) \\ &= \int_0^1 dh\, p(k'=1|n'=1,h)\, p(h|D) \\ &= \int_0^1 dh\, h\, p(h|D) \\ &= < h >_{\text{posterior}}, \end{aligned} \tag{2.132}$$

where k' represents the number of 'heads' in n' test observations. The multi–step predictive probability $p(k'|n',n,k)$, however, has not a form similar to Eq. (2.124) if expressed in terms of $< h >_{\text{posterior}}$. Indeed, denoting by k_m the outcome of the mth trial,

$$\begin{aligned} p(k'|n',n,k) &= p(k_{n+n'},\cdots,k_{n+1}|k_n,\cdots,k_1) \\ &= p(k_{n'}|k_{n'-1},\cdots,k_1)\cdots p(k_{n+1}|k_n,\cdots,k_1) \\ &= \int_0^1 dh\, p(k'|n',h)\, p(h|D) \\ &= \binom{n}{k} \int_0^1 dh\, h^k (1-h)^{(n-k)} p(h|D) \\ &\neq \binom{n}{k} < h >^k_{\text{posterior}} (1- < h >_{\text{posterior}})^{(n-k)}, \end{aligned} \tag{2.133}$$

as in general for nonlinear functions $< f(h) > \neq f(< h >)$. This is due to the fact that the test data itself contain information about h. For example, if $k = n$, Eq. (2.133) becomes $p(k'|n',n,k) = < h^n >_{\text{posterior}}$ which is larger or equal to $< h >^n_{\text{posterior}}$. That means if finding more and more 'heads' the probability to find another 'head' in the next trial grows (at least as long as the true h is not yet known). Note that it is thus not optimal to predict the next $n' > 1$ events at once. An optimal prediction proceeds step by step by calculating $p(k'|n'=1,n,k)$, taking into account always all the available data, including the already known 'test' data.

In cases where the integration over h to obtain the mean of the posterior (predictive probability) cannot be done exactly, we can try to find as appproximation of the mean the most probable h under the posterior $p(h|D)$, i.e., the mode of the posterior, also known as the Maximum A Posteriori (MAP) solution. Approximating the mean by the mode means replacing integration over h by maximization over h. We study now the approximation of the posterior mean by the posterior mode in the coin example. To find the posterior mode or MAP solution h^*

$$h^* = p_{\text{map}}(k'=1|n'=1) = p_{\text{map}}(\text{'head'}), \tag{2.134}$$

we have (for differentiable posteriors) to solve the stationarity equation

$$\frac{dp(h|D)}{dh} = 0 \tag{2.135}$$

or, equivalent but sometimes more convenient, $d\ln p(h|D)/dh = 0$.

For the uniform prior, the log–posterior (which is up to an arbitrary constant equal to the negative energy $-E(h)$) reads

$$\ln p(D|h) = k \ln h + (n-k)\ln(1-h) + \ln \binom{n}{k}. \tag{2.136}$$

The corresponding stationarity equation

$$0 = \frac{d}{dh}\ln p(h|D) = \frac{d}{dh}\ln p(D|h) = \frac{\partial \ln p(k|n,h)}{\partial h} = \frac{k}{n} - \frac{n-k}{1-h} \tag{2.137}$$

has the solution

$$h^* = p_{\text{map}}(\text{'head'}) = \frac{k}{n}, \tag{2.138}$$

which is just the observed relative frequency.

Similarly, for Gaussian prior the logarithm of the prior reads

$$\ln p(h) = -\frac{(h-h_0)^2}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right), \tag{2.139}$$

so that setting the derivative of the log–posterior with respect to h yields

$$0 = \frac{d}{dh}\ln p(h|D) = \frac{d}{dh}\ln p(D|h) + \frac{d}{dh}\ln p(h) = \frac{k}{h} - \frac{n-k}{1-h} - \frac{h-h_0}{\sigma^2}. \tag{2.140}$$

The resulting stationarity equation is in that case nonlinear

$$h^* = \frac{k}{n} - \frac{1}{n\sigma^2}(h-h_0)\,h\,(1-h). \tag{2.141}$$

Note that the presence of the prior pushes the MAP–solution from the observed relative frequency towards the prior mode h_0. In the limit $\sigma^2 \to \infty$ the second term on the right hand side vanishes so h^* approaches the observed relative frequency, i.e., the solution for a uniform prior. For vanishing width of the Gaussian, $\sigma^2 \to 0$, the solution $h = h_0$ becomes equal to the mode of the prior. (For $h_0 \neq 0$ or $h_0 \neq 1$ the two other solutions, $h = 0$ and $h = 1$, of Eq. (2.141) in that limit $\sigma^2 \to \infty$ represent a minimum and not a maximum of the posterior.)

Figs. 2.10, 2.11, 2.12, 2.13, 2.14, 2.15 show data, likelihood, prior, posterior, posterior mean (p_{pred}), and mode (p_{map}) for several different values of n and different priors. Fig. 2.16 and Fig. 2.17 show, for varying k and varying n, how different posteriors can be for the three different priors.

A quantitative measure of the accuracy of the estimation $< h >_{\text{posterior}}$ is the standard deviation of the posterior (error bar)

$$\sigma = \sqrt{< h^2 >_{\text{posterior}} - < h >^2_{\text{posterior}}}. \tag{2.142}$$

As an approximation of the true error bar one can calculate the error bar σ_{map} in Maximum A Posteriori Approximation which, as in Eq. (2.93), is defined as the standard deviation of the Gaussian approximation at h^* according to

$$e^{-f(x)} \approx e^{-f(x^*) - \frac{(x-x^*)^2}{2} \frac{d^2 f(x)}{dx^2}\Big|_{x=x^*}} = e^{-f(x^*)} e^{-\frac{(x-x^*)^2}{2\sigma^2_{\text{map}}}}. \tag{2.143}$$

To find the MAP error bar σ_{map} for the binomial model with uniform prior we have to calculate

$$0 = -\frac{d^2}{dh^2} \ln p(D|h) = -\frac{k}{h^2} + \frac{n-k}{(1-h)^2} \tag{2.144}$$

which becomes at $h = h^* = \frac{k}{n}$,

$$0 = -\frac{n^2}{k} + \frac{n-k}{\left(1-\frac{k}{n}\right)^2} = \frac{n}{h^*(1-h^*)}. \tag{2.145}$$

Thus, we have as MAP error bar

$$\sigma_{\text{map}} = \sqrt{\frac{h^*(1-h^*)}{n}}. \tag{2.146}$$

Note, that the error bar vanishes with $1/\sqrt{n}$ for large n.

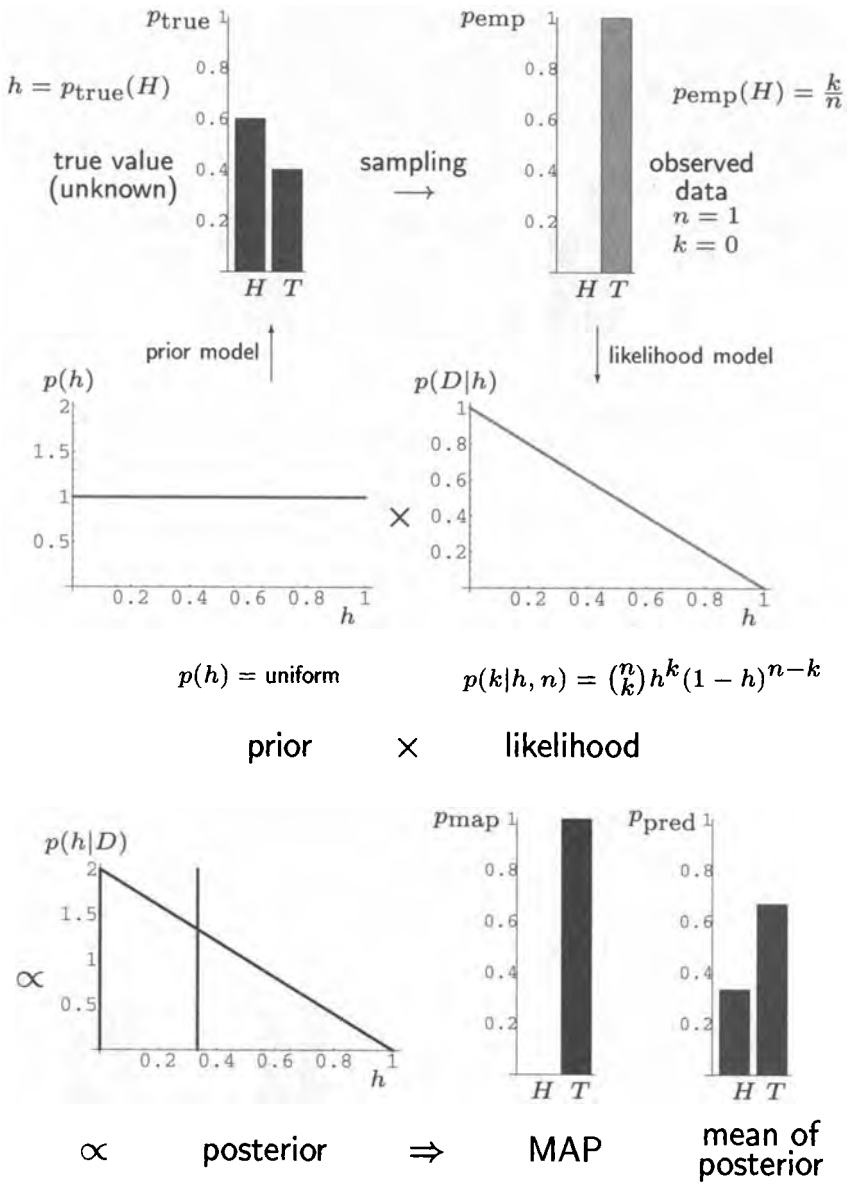


Figure 2.10: The coin example 1 (uniform prior, $n = 1$). Shown is the calculation of the posterior for a coin (binary event with H = 'head' and T = 'tail') with uniform prior (2.126), 1 data point ($n = 1$) having found 'tail' ($k = 0$), and a true probability $h_{\text{true}} = p_{\text{true}}(H)$ of 0.6 to find 'head'. For the coin example the space of hypotheses ϕ is spanned by all possible probabilities to find 'head' tossing a coin once, $\phi = h = p(\text{'head'}) = p(H)$, $0 \leq h \leq 1$. The data $D = (k, n)$, tossing a coin n–times, are represented by the observed number k of 'heads' sampled randomly from a binomial likelihood $p(D|h) = p(k|n,h) = \binom{n}{k} h^k (1-h)^{(n-k)}$.

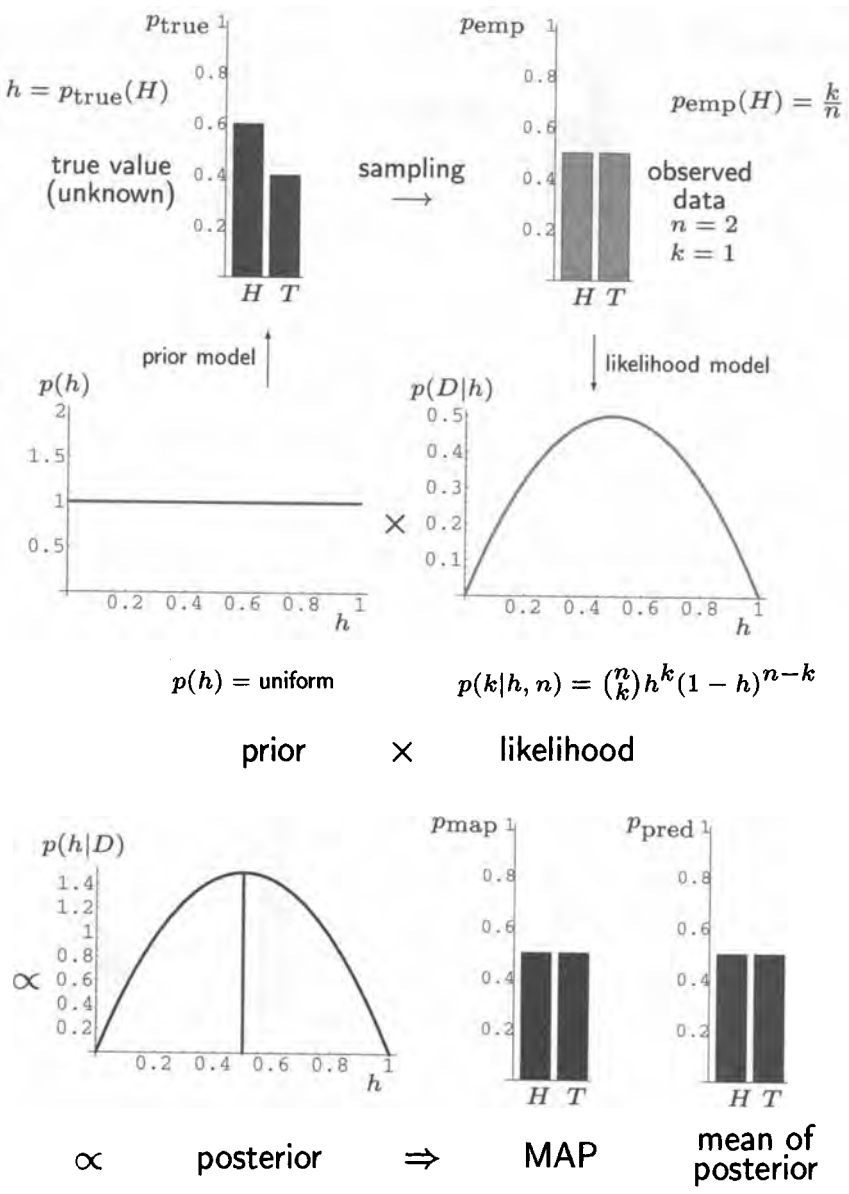


Figure 2.11: The coin example 2 (uniform prior, $n = 2$). Like Fig. 2.10 but with 2 data points ($n = 2$) having found one times 'tail' and one times 'head' ($k = 1$).

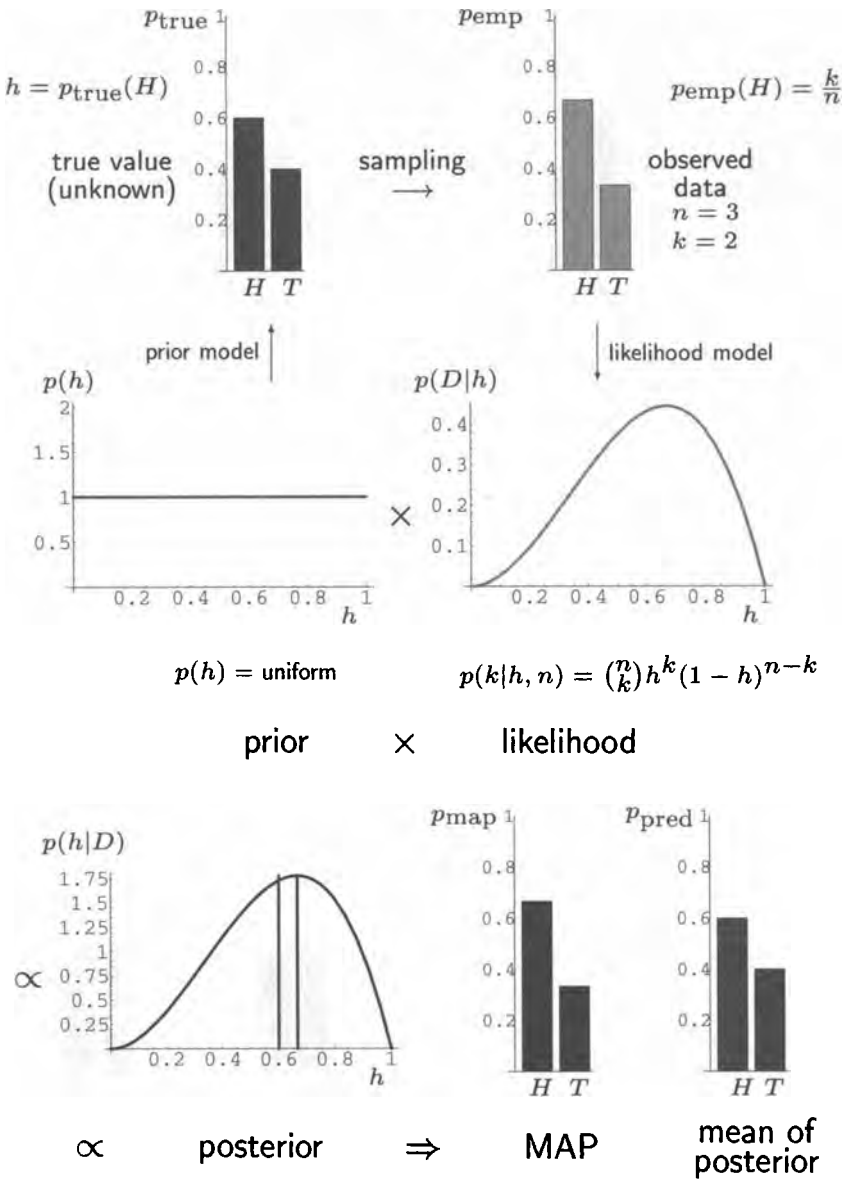


Figure 2.12: The coin example 3 (uniform prior, $n = 3$). Like Fig. 2.10 but with 3 data points ($n = 3$) having found one times 'tail' and two times 'head' ($k = 2$).

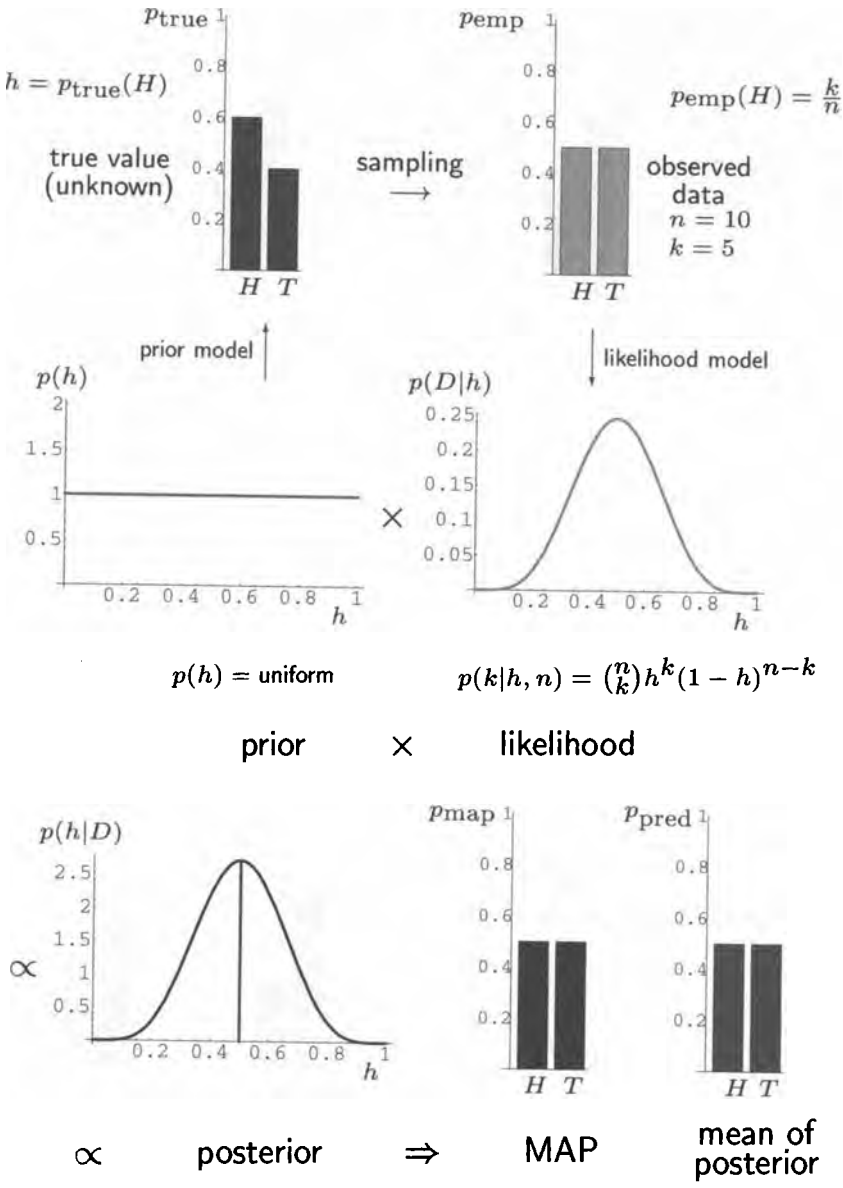


Figure 2.13: The coin example 4 (uniform prior, $n = 10$). Like Fig. 2.10 but with 10 data points ($n = 10$) having found five times 'tail' and five times 'head' ($k = 5$).

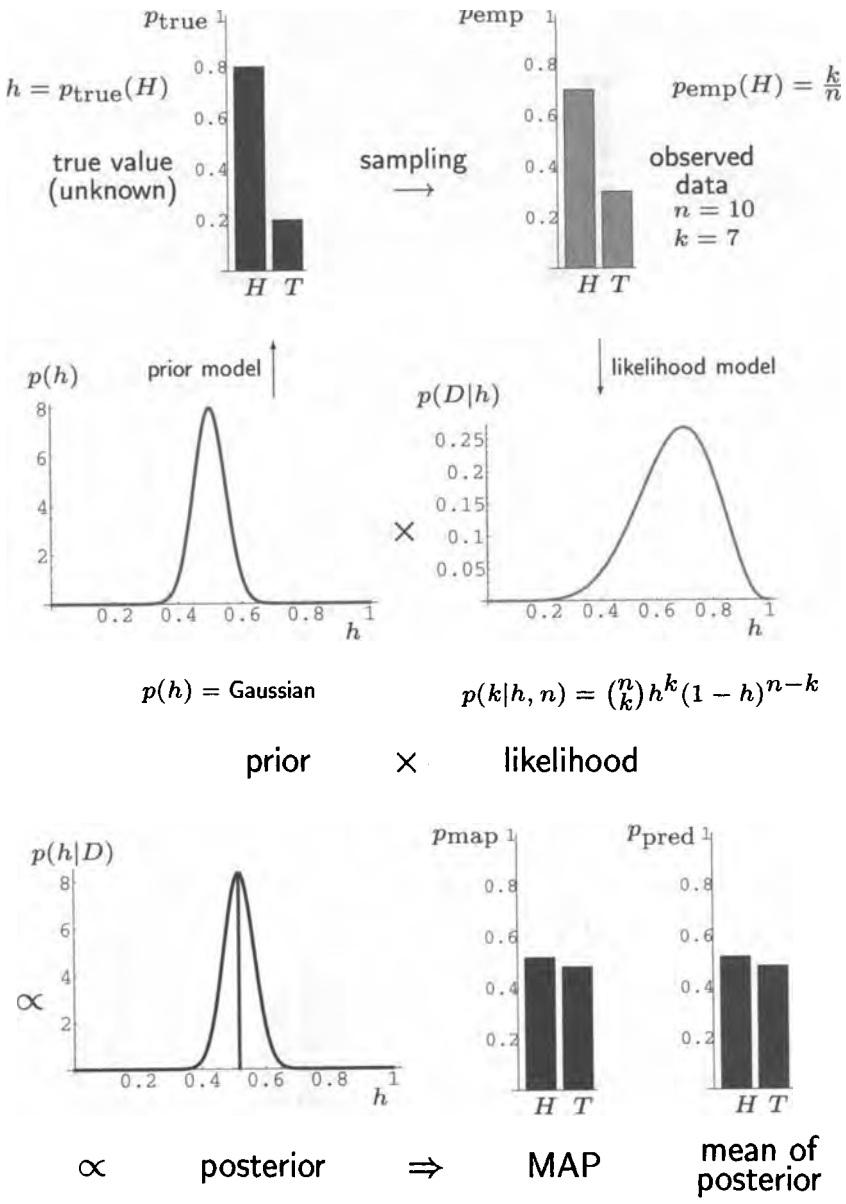


Figure 2.14: The coin example 5 ('expert' prior, $n = 10$). Like Fig. 2.10 but with a (nearly) Gaussian prior sharply peaked around 0.5 ('expert' prior, (2.128)), 10 data points ($n = 10$) having found three times 'tail' and seven times 'head' ($k = 5$), and a true probability $h_{\text{true}} = p_{\text{true}}(H)$ of 0.8 to find 'head'.

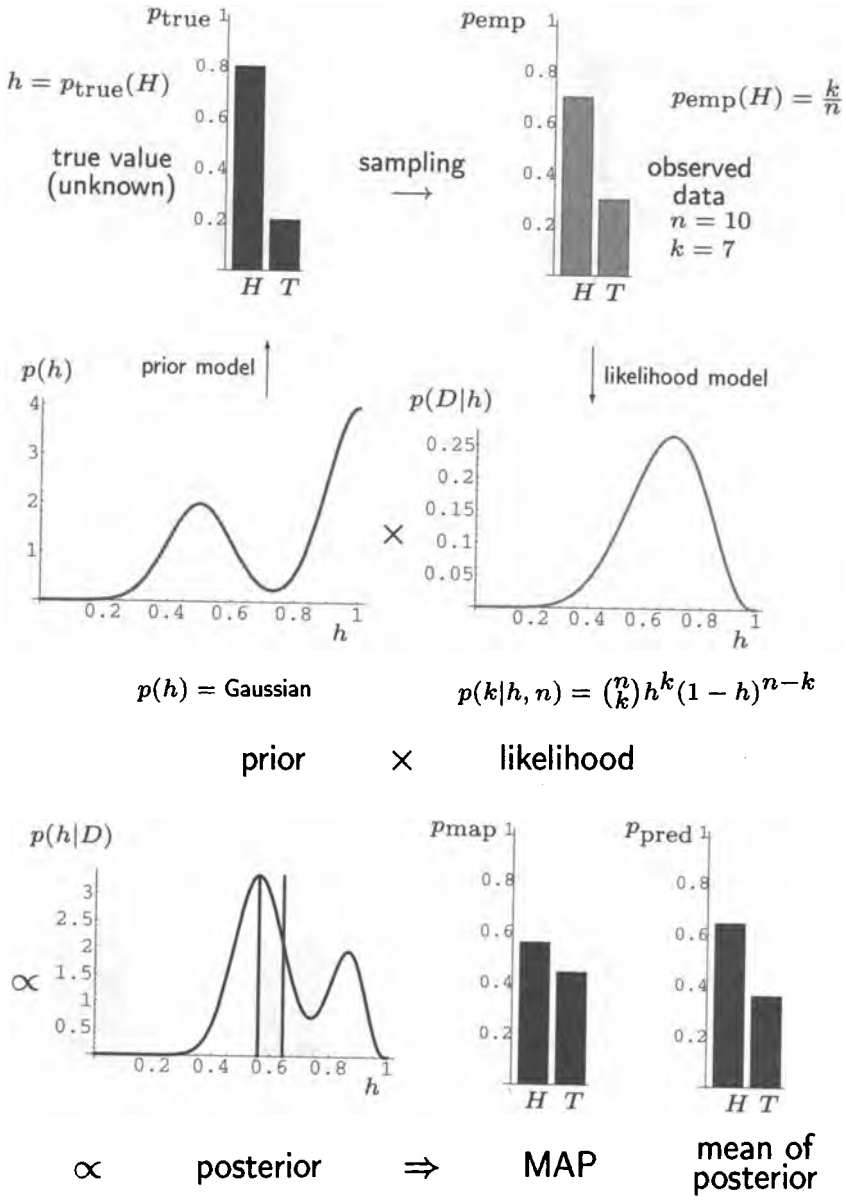


Figure 2.15: The coin example 6 ('Las Vegas' prior, $n = 10$). Like Fig. 2.10 but with 'Las Vegas' prior (2.129), 10 data points ($n = 10$) having found three times 'tail' and seven times 'head' ($k = 7$), and a true probability $h_{\text{true}} = p_{\text{true}}(H)$ of 0.8 to find 'head'.

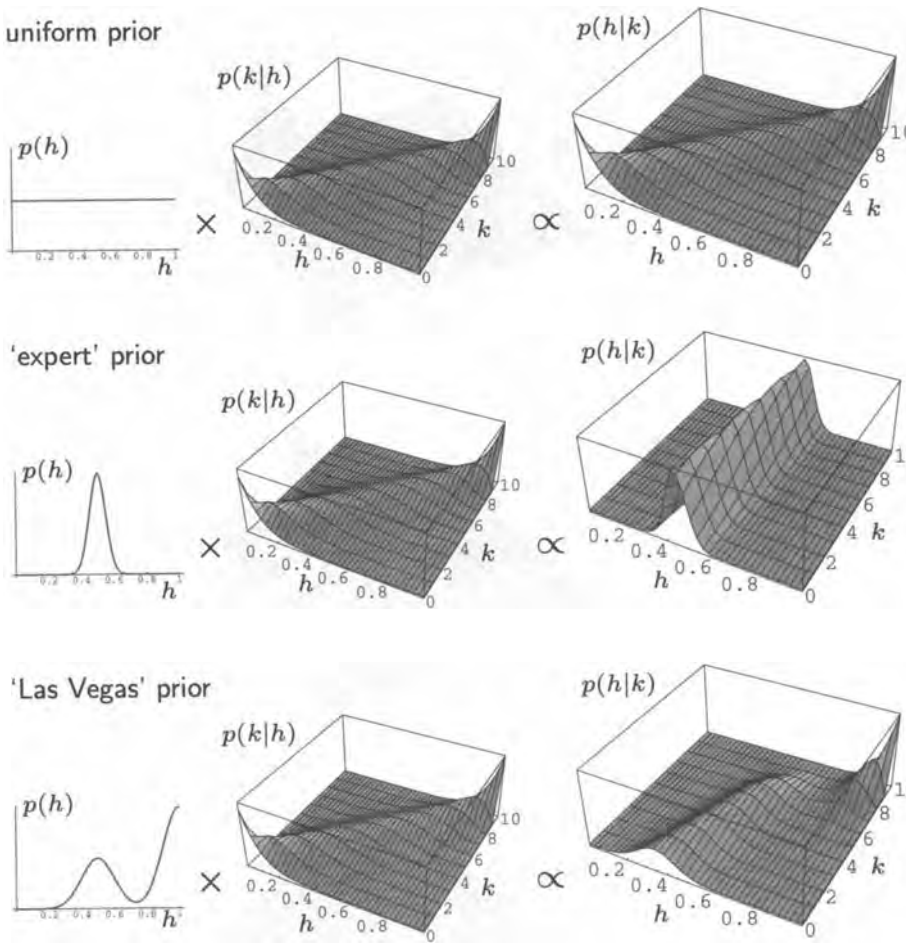


Figure 2.16: The coin example 7 (posteriors for three different priors). Shown is the dependency of the posterior $p(h|D) = p(h|n,k)$ from the number k of finding 'head' tossing a coin $n = 10$ times for the three different priors (2.126, 2.128, 2.129).

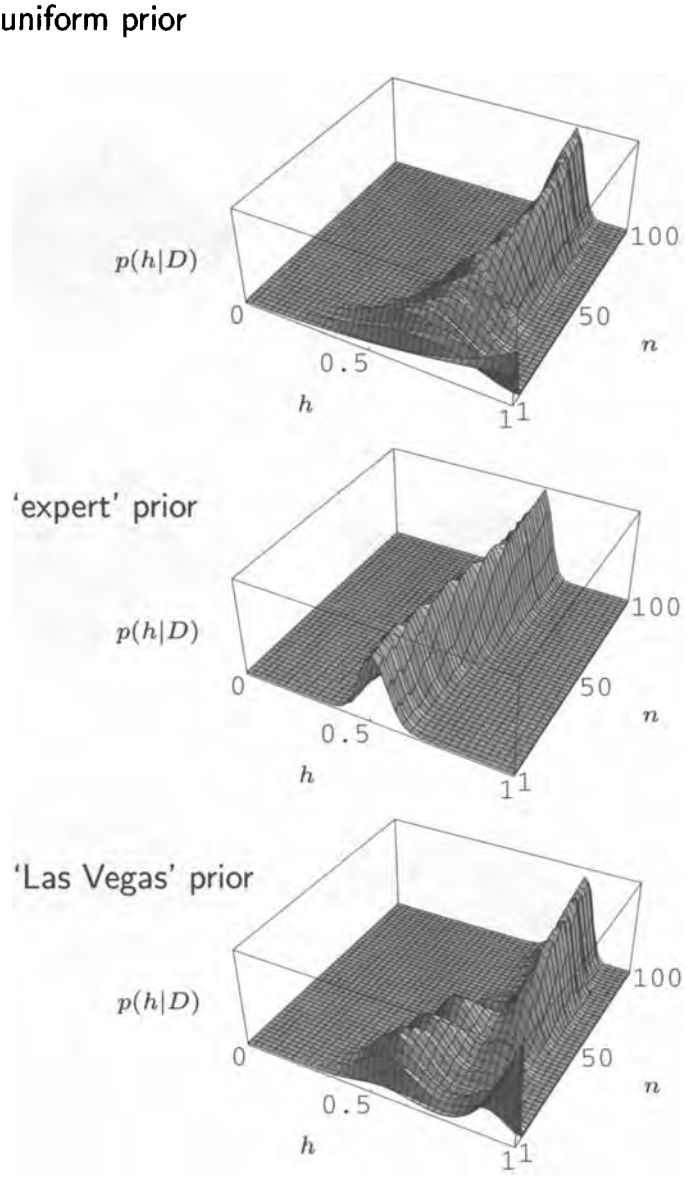


Figure 2.17: The coin example 8 (evolution of the three posteriors). A fair coin? Shown are the resulting posteriors for the three different priors (2.126, 2.128, 2.129) while sampling $n = 100$ random data points for true $h_{\text{true}} = 0.8$, using the same random data for the three cases. Note that different priors result in quite different conclusions about the coin even for identical data.

2.2 Bayesian decision theory

2.2.1 Loss and risk

Bayesian learning means making predictions on the basis of data and *a priori* information. Hence learning constitutes the decision problem of choosing one specific prediction out of a set of possible predictions. Predicting correct or false values typically has consequences in practice. *Bayesian decision theory* provides a general framework to include these aspects.[5]

In Bayesian decision theory a set A of possible actions a is considered, together with a loss function $l(x, y, a)$ describing the *loss* suffered in situation x if y appears and action a is selected [33,285,386,424,438]. The loss averaged over *test data* x, y is known as *expected risk*,

$$r(a, f) = \int dx\, dy\, p(x)\, p(y|x, f)\, l(x, y, a). \tag{2.147}$$

$$= < l(x, y, a) >_{X,Y|f} \tag{2.148}$$

where $< \cdots >_{X,Y|f}$ denotes the expectation under the joint predictive density $p(x, y|f) = p(x)p(y|x, f)$ of x and y for a given state of knowledge $f = \{D, D_0, D_S\}$ (see Section 2.1.1). Expressing $p(y|x, f)$ according to Eq. (2.64) as an integral over ϕ (representing the possible states of Nature), the expected risk can also be written

$$r(a, f) = \int dx\, dy\, d\phi\, p(x)\, p(y|x, \phi) p(\phi|D, D_0)\, l(x, y, a). \tag{2.149}$$

$$= < r(a, \phi) >_{\Phi|f} \tag{2.150}$$

where

$$r(a, \phi) = \int dx\, dy\, p(x)\, p(y|x, \phi)\, l(x, y, a), \tag{2.151}$$

and $< \cdots >_{\Phi|f}$ denotes the expectation under the posterior $p(\phi|D, D_0)$. The aim is to find an optimal action

$$a^* = \text{argmin}_{a \in A} r(a, f). \tag{2.152}$$

2.2.2 Loss functions for approximation

Log–loss: A typical loss function for *density estimation problems* is the *log–loss*

$$l(x, y, a) = -b_1(x) \ln p(y|x, a) + b_2(x, y) \tag{2.153}$$

[5]Bayesian decision theory is a normative theory [504]. Decisions of individuals in real life typically from optimal Bayesian decisions. Kahnemann and Tversky analysed heuristics and common biases which are often found in real decisions [156,248–250,288,487,488]. For example, they describe in their 'prospect theory' the dependence of psychological preferences on a 'reference point'. For the 'unfair coin' which we have treated in the last section, the effects of the so–called 'conservatism' and 'representativeness bias' in the assigment of posterior probabilities are, for example, discussed in [454] (p. 129).

with some a–independent $b_1(x) > 0$, $b_2(x,y)$ and actions a corresponding to probability densities $p(y|x,a)$ with

$$\int dy\, p(y|x,a) = 1, \ \forall x \in X, \forall a \in A. \tag{2.154}$$

Choosing $b_2(x,y) = \ln p(y|x,f)$ and $b_1(x) = 1$ gives

$$\begin{aligned} r(a,f) &= \int dx\, dy\, p(x)p(y|x,f) \ln \frac{p(y|x,f)}{p(y|x,a)} && (2.155) \\ &= < \ln \frac{p(y|x,f)}{p(y|x,a)} >_{X,Y|f} && (2.156) \\ &= < \mathrm{KL}(\, p(y|x,f),\, p(y|x,a)\,) >_X, && (2.157) \end{aligned}$$

which shows that minimizing log–loss is equivalent to minimizing the (x–averaged, continuous) *Kullback–Leibler entropy* $\mathrm{KL}(\, p(y|x,f),\, p(y|x,a)\,)$ [29, 103, 115, 275, 276]. (This continuous version of the discrete Kullback–Leibler entropy (2.36), $KL(p,q) = \sum_x p(x) \ln(p(x)/q(x))$, also called Kullback–Leibler distance and cross entropy, already appeared in Eq. (2.47) as relative entropy. The mutual information in Eq. (2.70) can be regarded as the Kullback–Leibler distance between the joint $p(x,y)$ and the factorial $p(x)p(y)$.) In case the family of possible actions $p(y|x,a)$ includes the full Bayesian predictive density $p(y|x,f)$ the Kullback–Leibler entropy is minimized by $p(y|x,a^*) = p(y|x,f)$ (see Section 2.2.4).

While this book will concentrate on log–loss, we will also give a short summary of loss functions for *regression problems*. (See [33, 424] for details.) Regression problems are special density estimation problems where the possible actions are restricted to y–independent functions $a(x)$.

Squared–error loss: The most common loss function for regression problems (see Section 3.8) is the squared–error loss. It reads for one–dimensional y

$$l(x,y,a) = b_1(x)\,(y - a(x))^2 + b_2(x,y), \tag{2.158}$$

with arbitrary $b_1(x) > 0$ and $b_2(x,y)$. In that case the optimal function $a(x)$ is the *regression function* of the posterior which is the mean of the predictive density

$$a^*(x) = \int dy\, y\, p(y|x,f) = < y >_{Y|x,f}\,. \tag{2.159}$$

This can be easily seen by writing

$$\begin{aligned} (y - a(x))^2 &= (y - < y >_{Y|x,f} + < y >_{Y|x,f} - a(x))^2 && (2.160) \\ &= (y - < y >_{Y|x,f})^2 + (a(x) - < y >_{Y|x,f})^2 \\ &\quad -2\,(y - < y >_{Y|x,f})\,(a(x) - < y >_{Y|x,f})\,, && (2.161) \end{aligned}$$

where the first term in (2.161) is independent of a and the last term vanishes after averaging over y with $p(y|x,f)$ according to the definition of $< y >_{Y|x,f}$. Hence

$$r(a,f) = \int dx\, b_1(x)p(x)\left(a(x) - < y >_{Y|x,f}\right)^2 + \text{const.} \tag{2.162}$$

This is minimized by $a(x) =< y >_{Y|x,f}$. Notice that, for Gaussian $p(y|x,a)$ with fixed variance, log–loss and squared–error loss are equivalent. Loss functions depending explicitly on multidimensional y can be defined. For instance, a general quadratic loss function would be

$$l(x,y,a) = \sum_{k,k'}(y_k - a_k(x))\mathbf{K}(k,k')(y_{k'} - a_{k'}(x)). \tag{2.163}$$

with symmetric, positive definite kernel $\mathbf{K}(k,k')$.

Absolute loss: For absolute loss

$$l(x,y,a) = b_1(x)|y - a(x)| + b_2(x,y) \tag{2.164}$$

with arbitrary $b_1(x) > 0$ and $b_2(x,y)$, the risk becomes

$$\begin{aligned} r(a,f) &= \int dx\, b_1(x)p(x)\int_{-\infty}^{a(x)} dy\,(a(x) - y)\,p(y|x,f) \\ &\quad + \int dx\, b_1(x)p(x)\int_{a(x)}^{\infty} dy\,(y - a(x))\,p(y|x,f) + \text{const.} \\ &= 2\int dx\, b_1(x)p(x)\int_{m(x)}^{a(x)} dy\,(a(x) - y)\,p(y|x,f) + \text{const.}', \end{aligned} \tag{2.165}$$

where the integrals have been rewritten as $\int_{-\infty}^{a(x)} = \int_{-\infty}^{m(x)} + \int_{m(x)}^{a(x)}$ and $\int_{a(x)}^{\infty} = \int_{a(x)}^{m(x)} + \int_{m(x)}^{\infty}$, introducing a median function[6] $m(x)$ which satisfies

$$\int_{-\infty}^{m(x)} dy\, p(y|x,f) = \frac{1}{2},\ \forall x \in X, \tag{2.166}$$

and we used

$$a(x)\left(\int_{-\infty}^{m(x)} dy\, p(y|x,f) - \int_{m(x)}^{\infty} dy\, p(y|x,f)\right) = 0,\ \forall x \in X, \tag{2.167}$$

to obtain Eq. (2.165). The risk (2.165) is minimized by any *median function* $m(x)$ where the integral vanishes. Thus the optimal approximation $a(x) = m(x)$ is for each x a median of the predictive density $p(y|x,f)$.

[6]Because the median is not necessarily unique, there can exist more than one median function.

δ–loss and 0–1 loss: Another possible loss function, typical for classification tasks (see Section 3.9) like image segmentation [332], is the δ–loss for continuous y or 0–1–loss for discrete y

$$l(x, y, a) = -b_1(x)\delta(y - a(x)) + b_2(x, y), \tag{2.168}$$

with arbitrary $b_1(x) > 0$ and $b_2(x, y)$. Here δ denotes the Dirac δ–functional for continuous y and the Kronecker δ for discrete y. The corresponding risk

$$r(a, f) = -\int dx\, b_1(x) p(x)\, p(\, y = a(x) \,|x, f) + \text{const.} \tag{2.169}$$

is maximized if for every x the $a(x) = y$ with maximal $p(y|x, f)$ is chosen. Thus the optimal approximation $a(x)$ is a *mode function* of the predictive density $p(y|x, f)$.[7] For Gaussians, mode and median are unique and coincide with the mean.

2.2.3 General loss functions and unsupervised learning

Choosing actions a in specific situations often requires the use of specific loss functions. Such loss functions may, for example, contain additional terms measuring *costs* of choosing action a not related to the quality of the approximation. Such costs can quantify aspects like the simplicity, implementability, production costs, sparsity, or understandability of action a.

Instead of approximating a whole density it often suffices to extract some of its features, like identifying clusters of similar y–values, finding independent components for multidimensional y, or mapping to an approximating density with lower dimensional x. This kind of *exploratory data analysis* is the Bayesian analogue to *unsupervised learning methods* [209, 269, 270, 419]. Such methods are also often utilized as a preprocessing step.

From a Bayesian point of view general loss functions require an explicit two–step procedure [291]:

1. calculate (an approximation of) the predictive density, and
2. minimize the expectation of the loss function under that (approximate) predictive density.

For an example see [301]. Empirical risk minimization discussed in Section 2.2.5, on the other hand, minimizes the empirical average of the (possibly regularized) loss function.

For a Bayesian version of cluster analysis, i.e., the partitioning of a predictive density obtained from empirical data into several clusters, a possible loss function is

$$l(x, y, a) = (y - a(x, y))^2, \tag{2.170}$$

[7] Like a median, a mode (location(s) where p is maximal) is not necessarily unique, so there can exist more than one mode function.

with action $a(x,y)$ being a mapping of y for given x to a finite number of cluster centers (prototypes). Another example of a clustering method based on the predictive density is Fukunaga's valley seeking procedure [140].

For multidimensional x it can be useful to choose a space of actions $a(\mathbf{P}_x x, y)$ depending only on a (possibly adaptable) lower dimensional projection $\mathbf{P}_x$ of x. For multidimensional y with components y_i it is often useful to identify independent components. One can look for a linear mapping $y = \mathbf{M}\tilde{y}$ minimizing the correlations between different components of the 'source' variables $\tilde{y}$ by minimizing a loss function including a term like $l(y,y',\mathbf{M}) = \sum_{i\neq j} \tilde{y}_i \tilde{y}'_j$ with respect to $\mathbf{M}$ under the joint predictive density for y and y', given x, x', D, D_0. This results in a Bayesian version of blind source separation (cf. the so–called cocktail party problem [16, 30]), analogous to the treatment of Molgedey and Schuster [349]. Interesting projections of multidimensional y can be found by projection pursuit techniques (see Section 4.9 and [139, 227, 243, 456]).

2.2.4 Log–loss and Maximum A Posteriori Approximation

In Section 2.1.6 we have discussed the Maximum A Posteriori Approximation for evaluating the integral over all states ϕ in the predictive density (2.64). Interestingly, the stationary ϕ^* with maximal posterior $p(\phi|f)$ is not only useful to obtain an approximation for the predictive density $p(y|x,f)$, but is also the optimal solution a^* for a Bayesian decision problem with log–loss and $a \in A = \Phi$. Similarly, if the set of possible actions $p(y|x,a)$ includes the full Bayesian predictive density $p(y|x,f) \in F$ then $p(y|x,a^*) = p(y|x,f)$ is the optimal solution. Indeed, for a Bayesian decision problem with log–loss (2.153)

$$a^* = \text{argmin}_{a\in\Phi} r(a,\phi) = \phi, \tag{2.171}$$

i.e., $p(y|x,a^*) = p(y|x,\phi)$. Analogously

$$a^* = \text{argmin}_{a\in F} r(a,f) = f, \tag{2.172}$$

i.e., $p(y|x,a^*) = p(y|x,f)$. For log–loss solving the approximation problem for the predictive density $p(y|x,f)$ already solves the full decision problem.

This is proved as follows: Jensen's inequality states that

$$\int dy\, p(y) g(q(y)) \geq g(\int dy\, p(y) q(y)), \tag{2.173}$$

for any convex function g and probability density $p(y) \geq 0$ with $\int dy\, p(y) = 1$. Thus, because the logarithm is concave, the Kullback–Leibler entropy (2.155) fulfills

$$-\int dy\, p(y|x,\phi) \ln \frac{p(y|x,a)}{p(y|x,\phi)} \geq -\ln \int dy\, p(y|x,\phi) \frac{p(y|x,a)}{p(y|x,\phi)} = 0 \tag{2.174}$$

$$\Rightarrow -\int dy\, p(y|x,\phi) \ln p(y|x,a) \geq -\int dy\, p(y|x,\phi) \ln p(y|x,\phi), \qquad (2.175)$$

with equality for $a = \phi$. Hence

$$\begin{aligned} r(a,\phi) &= -\int dx \int dy\, p(x)p(y|x,\phi)\,(b_1(x) \ln p(y|x,a) - b_2(x,y)) \\ &= -\int dx\, p(x)b_1(x) \int dy\, p(y|x,\phi) \ln p(y|x,a) + \text{const.} \quad (2.176) \\ &\geq -\int dx\, p(x)b_1(x) \int dy\, p(y|x,\phi) \ln p(y|x,\phi) + \text{const.} \quad (2.177) \\ &= r(\phi,\phi), \end{aligned}$$

with equality for $a = \phi$. For $a \in F$ replace $\phi \in \Phi$ by $f \in F$. This proves Eqs. (2.171) and (2.172).

2.2.5 Empirical risk minimization

In the previous sections the (error or energy) functionals to be minimized in MAP have been given a Bayesian interpretation in terms of the log–posterior density. The Frequentist approach of (regularized) *empirical risk minimization* [492–496] provides an alternative way to construct error functionals as a (regularized) plug–in estimate for the expected loss.[8]

The Frequentist approaches to statistics are based on the interpretation of probabilities as (asymptotic) relative empirical frequencies (see also Appendix A). Hidden variables ϕ (unknown parameters, theories, models) which can not be sampled empirically should appear only as conditioning variables, as in the likelihood $p(y|x,\phi)$, but not as random variables as in a prior $p(\phi)$. We will now sketch shortly the three main Frequentist approaches which are,

1. *hypothesis testing*,
2. the *maximum likelihood approach*, which is technically equivalent to a MAP with uniform prior, and
3. a *regularized empirical risk minimization*, where the error functional for training data is based on the loss function instead of the likelihood and a regularizer plays the role of the prior.

For simple statistical questions, the standard Frequentist approach is *hypothesis testing*, where the data are analyzed for one selected 'null' hypothesis ϕ_0. For example, being interested in the effect of some (e.g., medical) treatment, ϕ_0 typically represents the hypothesis that the treatment shows no effect.

[8] Prior-like terms in error functionals have also been interpreted as measure of complexity where functions with high probability correspond to functions of small complexity [162,399, 417,418,494].

The statistical decision of 'accepting' or 'rejecting' the null hypothesis is then based on the P-value, being the probability that equal or more extreme data are observed if the selected null hypothesis would be true. For example, in one–sided testing having measured n independent data $D = (x_i, y_i | 1 \leq i \leq n)$ the P-value is $\int_{y_D}^{\infty} dy'_D \, p(y'_D|x_D, \phi_0) = P(y'_D \geq y_D|\phi_0)$ with $p(y'_D|x_D, \phi_0) := \prod_i p(y'_i|x_i, \phi_0)$, x_D denoting the vector of x_i, y_D the vector of y_i, and y'_D the vector of integration variables. The null hypothesis is rejected if the P–value is small enough (significant), say smaller than 1% or 5%.

Thus while in a full Bayesian approach the data are fixed and one integrates over hypotheses ϕ, in hypothesis testing a hypothesis ϕ_0 is fixed and one integrates over data (see also Figs. 2.18 and 2.19). Due to the factorization the integration over (independent) data is typically easier to perform than the integration over hypotheses, especially if the hypothesis space is large. Some problems related with hypothesis testing are [33,61,238]:

1. One 'natural' null hypothesis must be selected.
2. It must be defined what 'more extreme' means.
3. The P–value can *not* be interpreted as probability that the null hypothesis is true [238,310,438].
4. Rejecting a null hypothesis does not result in a specific alternative point hypothesis.
5. Sampling enough data, even arbitrarily small deviations from the null hypothesis lead to its rejection.
6. Testing multiple hypotheses using the P–values for single hypotheses increases the chance of finding significant results. Simultaneous testing of multiple hypotheses requires the P–value to be adapted (see [199,219, 342,437]). (This is related to the problem that significant results are probably more likely to get published than non–significant findings.)
7. The integration over data includes data which have not been observed. This makes P–values dependent not only on the observed data but also on 'possible data', for example, the experimental design, and violates the 'likelihood principle' [33,312]. Observing 9 times 'head' and 3 times 'tail' tossing a coin, Frequentist P–values differ if the experiment was designed to stop after tossing the coin 12 times (binomial likelihood) or after having found 'tail' 3 times (negative binomial likelihood). In contrast Bayesian approaches depend only on observed data (and the prior model) but not on such 'stopping rules'. (See Fig. 2.20.)

For more complex parametric models the classical Frequentist method is the *maximum likelihood approach*. Technically, this is equivalent to a Bayesian MAP with uniform prior. If the number of parameters is too large to find a

well defined maximum likelihood solution (like for nonparametric models), additional penalty or regularization terms are added. This method, being technically a *regularization approach*, is also called *penalized maximum likelihood approach.*

Alternatively, error functionals to be minimized can be derived from a decision theoretic point of view. Like Bayesian decision theory the corresponding Frequentist approach of empirical risk minimization aims to minimize the *expected risk* for action a under the unknown true state of Nature ϕ_{true}

$$r(a, f) = \int dx\, dy\, p(x, y|\phi_{\text{true}})\, l(x, y, a). \tag{2.178}$$

The expected risk, however, can not be calculated without knowledge of the true state ϕ_{true}. In contrast to the Bayesian approach of introducing a model and working with a model–dependent $p(x, y|f(D, D_0))$ to approximate $p(x, y|\phi_{\text{true}})$, the Frequentist approach approximates the expected risk by the *empirical risk*

$$E(a) = \hat{r}(a, f) = \sum_i l(x_i, y_i, a). \tag{2.179}$$

Thus the empirical risk is the plug–in estimate of the expected risk obtained by replacing the unknown true probability by an observed empirical probability (consisting of δ–peaks) [120]. To be able to derive results on asymptotic convergence it is then necessary to assume that training data are sampled according to the true $p(x, y|f) = p(y|x, f)p(x)$ [113, 254, 285, 403, 492, 494]. Notice that the Bayesian predictive density is conditional on the given data D and does according to Eq. (2.60) not depend on the density $p(x)$ for the training data. A Bayesian MAP calculation therefore does not rely on specific assumptions on the distribution of x. The $p(x)$ for the test, which appears in the Bayesian framework in the expected risk (2.148), can in a Bayesian approach thus be different from the $p(x)$ for the training data. For a discussion of the relation between quadratic error functionals and Gaussian processes see also [259, 260, 309, 332, 381, 383, 384, 507].

From the Frequentist point of view the data–dependent part of the loss function (also sometimes called 'discrepancy') is not necessarily related to the log–likelihood of the data as they arise from the posterior–related Bayesian interpretation. However, like in the Bayesian approach, training data terms are not enough to make the minimization problem well defined for nonparametric models. In such cases minimizing the risk corresponds to an inverse problem (in the general sense) [122, 221, 222, 262, 478, 492, 494] which can, according to the classical regularization approach [353, 482, 483], be treated by including additional *regularization (stabilizer) terms* in the loss function l. From the point of view of empirical risk minimization those regularization terms, which correspond to the prior terms in a Bayesian approach, have to be either con-

sidered as additional non–data related cost terms in the loss function or as a mere technical tool to make the minimization problem well defined.

The *empirical generalization error* for a test or validation data set, independent of the training data D, has then to be measured using only the data terms of the error functional without regularization terms. In empirical risk minimization this empirical generalization error is used, for example, to determine adaptive (hyper–)parameters of regularization terms. A typical example is a factor multiplying the regularization terms controlling the trade–off between data and regularization terms. Common Frequentist techniques to determine such parameters are *cross–validation* [12,89,169,356,471,472,507,509] (see also Section 5.4.6), *bootstrap* [118–120,193], or related techniques. Cross–validation or bootstrap methods also give directly an empirical guess for the stability (variance) of the parameters when sampling different training data. For Gaussian likelihoods (regression problems) a regularization parameter can also be determined using the *discrepancy method* self–consistent methods, or other methods which can be used if σ^2 is known [67,89,122,221,222,326,507]. The discrepancy method chooses the regularization parameter so that $\chi^2 = \sum_{i=1}^{n} |y_i - \phi(x_i)|/\sigma^2 = n$ (for known σ), which should be approximately fulfilled for large enough n. Equivalently one uses $\chi^2 = \sum_{i=1}^{n} |y_i - (\mathbf{F}\phi)(x_i)|/\sigma^2$ in a classical inverse problem setting where one is interested in the deconvolution of noise and averaging processes of the measurement device represented by $\mathbf{F}$ (e.g., removal of blurring). From a strict Bayesian point of view those hyperparameters have to be integrated out after defining an appropriate (hyper–)prior [33,52,317,323].

2.2.6 Interpretations of Occam's razor

The principle to prefer simple models over complex models and to find an optimal trade–off between fitting data and model complexity is often referred to as Occam's razor (William of Occam, 1285–1349) [60]. (For historical references see [215].) Regularization terms, penalizing, for example, non–smooth ('complex') functions, can be seen as an implementation of Occam's razor.

Similarly, a so–called Occam factor appears in Bayesian model selection resulting from the integration over model parameters calculated in Gaussian approximation [239,318,319,457] (cf. also [82] and the critique in [525]) or in low–temperature expansion including higher order terms [22]. Indeed, when studying the generalization behavior of trained models on a test set different from the training set, it is often found that there is an optimal model complexity. (See, for example, Figs. 3.16 – 3.27 in Section 3.8.5.) Complex models, due to their higher flexibility, can achieve better performance on the training data than simpler models. On a test set independent of the training set, however, they may perform poorer than simpler models. This phenomenon appearing in practical learning is known as 'overfitting' [52,209,492,494].

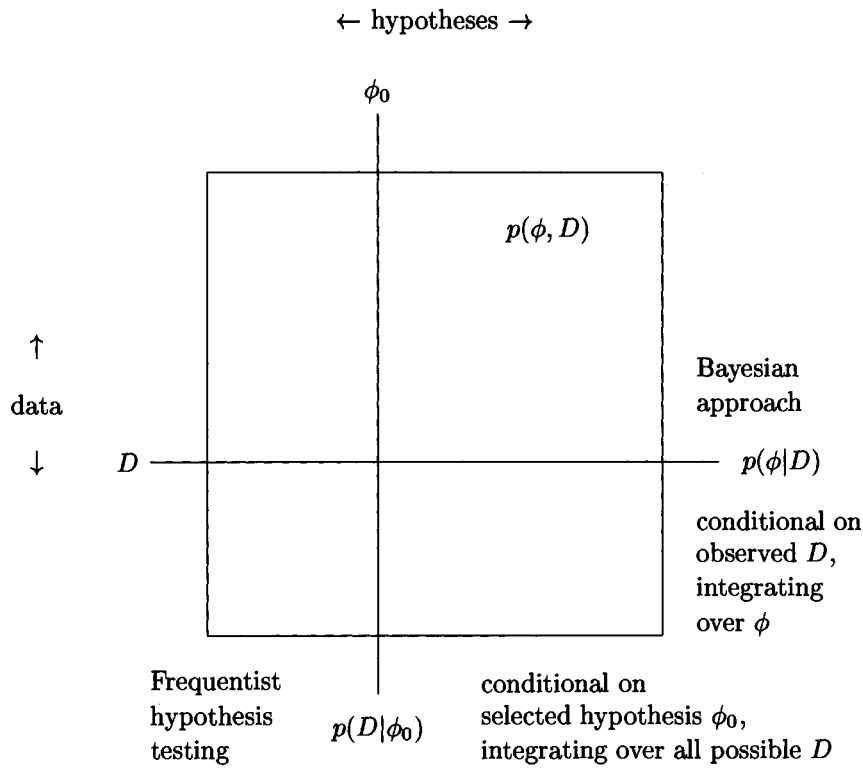


Figure 2.18: Bayesian approach vs. Frequentist hypothesis testing.

Notice, however, that the simplicity of a model has no absolute meaning but can only be defined relative to some reference model. 'Being constant', for instance, can only be interpreted with respect to some arbitrary base line or reference function (see also Appendix A). Hence referring to Occam's razor already requires that a reference model be defined. This is nothing else than *a priori* information. Hence Occam's razor can only be a heuristic help in practice if simplicity is 'intuitively' defined appropriate for the learning task. (The term 'intuitively' refers here to some implicit coupling between the model of the observer and the observed system, like the mechanism of *a posteriori* control discussed in Appendix A.) In general, however, referring to Occam's razor does not solve the principal problem of obtaining useful *a priori* information.

Despite the technical similarities between a Bayesian MAP and empirical risk minimization, it is important to remark that the Bayesian interpretation of regularization terms as *a priori* information about Nature and the Frequentist interpretation of regularization terms as additional cost terms in the

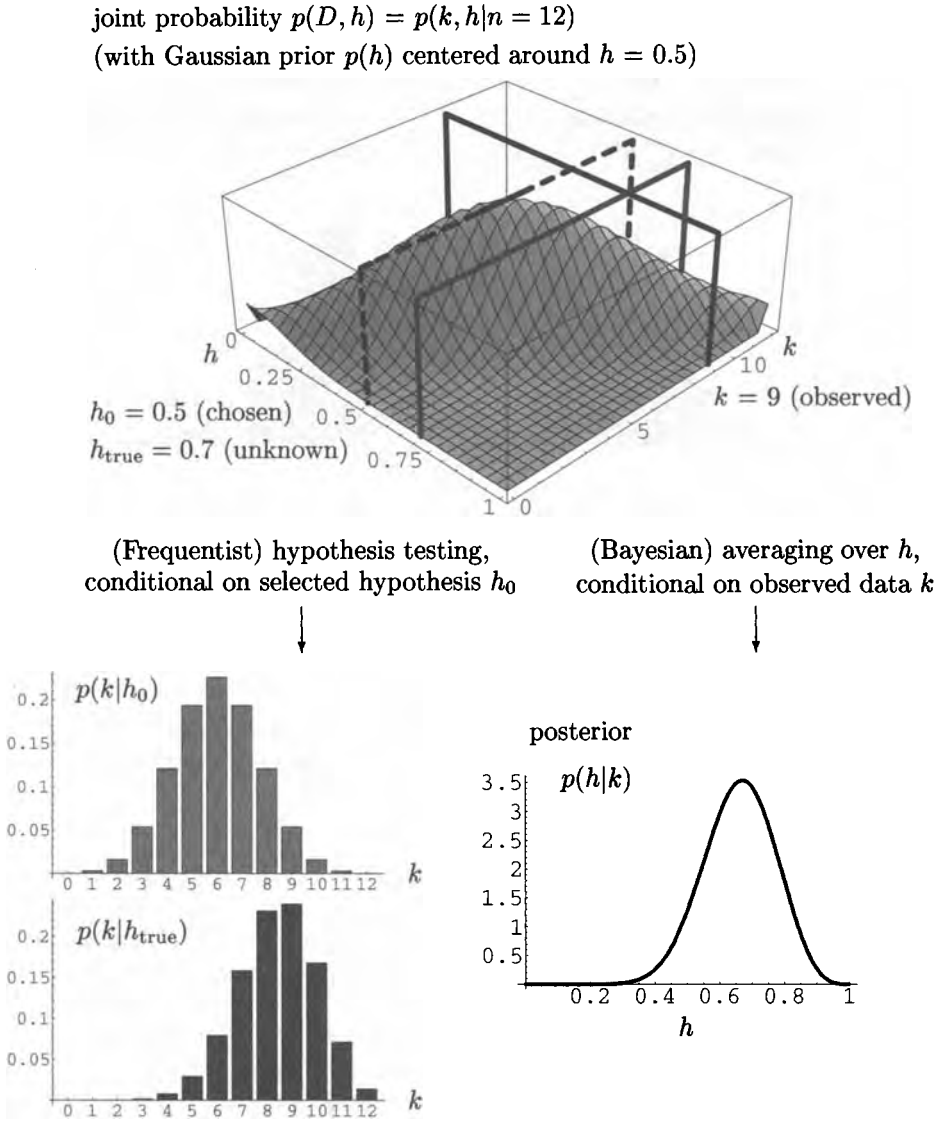
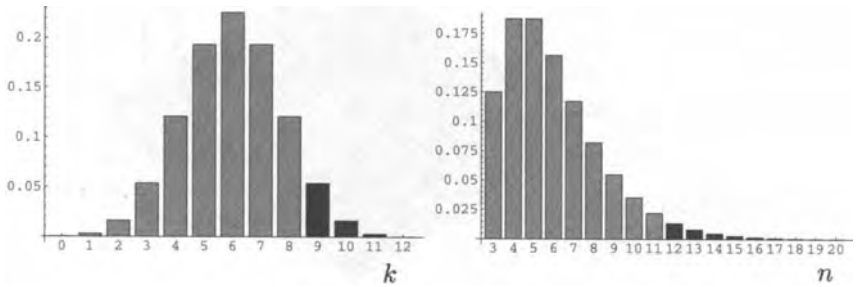


Figure 2.19: Frequentist hypothesis testing for null hypothesis $h_0 = 0.5$ vs. Bayesian approach with a Gaussian prior on $h = p(\text{'head'})$ for a coin (binomial likelihood) for $n = 12$ observations with $k = 9$ observed 'heads'. Increasing the number of data, the hypothesis testing approach will, in all cases where the null hypothesis h_0 is not (by accident) the exact true model, finally reject the null hypothesis (without giving a specific alternative). The Bayesian posterior will be more and more peaked around the true value $h = 0.7$.

Stop at $n = 12$
binomial distribution
$p(k|n,h) = \binom{n}{k} h^k (1-h)^{n-k}$

Stop at $n - k = 3$
negative binomial distribution
$p(n|k,h) = \binom{n-1}{n-k-1} h^k (1-h)^{n-k}$



$P(k \geq 9) = 0.073$
$= \sum_{k=9}^{n} p(k|n,h)$
not significant at 5%

$P(n \geq 12) = 0.033$
$= \sum_{n=12}^{\infty} p(n|k,h)$
significant at 5%

Figure 2.20: Example of the influence of stopping rules on P–values for a coin with $p(\text{'head'}) = h = 0.5$ after observing $k = 9$ times 'head' and $n - k = 3$ times 'tail' [33, 312]. As the binomial and negative binomial likelihood just differ by a constant factor (given the data), Bayesian approaches are not sensitive to such stopping rules.

loss function are *not* equivalent. Complexity priors reflect the case where Nature is known to be simple while complexity costs express the wish for simple models without the assumption of a simple Nature. Thus while the practical procedure of minimizing an error functional with regularization terms appears to be identical for empirical risk minimization and a Bayesian Maximum A Posteriori Approximation, the underlying interpretation for this procedure is different. We have seen in Section 2.2.4 that in the general case, i.e., for loss functions differing from log–loss, including prior and costs, requires from a Bayesian point of view to distinguish explicitly between model states ϕ and actions a. Even in saddle point approximation, this results in a two step procedure, where in a first step the hypothesis ϕ^* with maximal posterior is determined, while the second step minimizes the risk for action $a \in A$ under that hypothesis ϕ^* [291].

2.2.7 Approaches to empirical learning

Summarizing, we list the basic approaches to empirical learning:

1. General *Bayesian decision theory* allows to encompass aspects which go

beyond an approximation of likelihoods, like complexity costs. This general approach requires calculation of the likelihood for the given data, definition of a prior, application of Bayes' theorem to get the posterior, integration over model states to obtain the predictive density, and minimization of the resulting risk.

2. For log–loss, that means for pure approximation problems where the only aim is to approximate the unknown true likelihood $p(y|x,\phi)$ in an optimal way by $p(y|x,a)$ and no additional cost terms are included in the loss function, it is sufficient to calculate the predictive density. This *Bayesian approach for approximation problems* requires likelihood, prior, application of Bayes' theorem to find the posterior, and an integration over model states. This integration is typically done by Monte Carlo methods.

3. In *Maximum A Posteriori Approximation*[9], integration over model states is replaced by maximization of the posterior. This requires likelihood, prior, and application of Bayes' theorem to get the posterior to be maximized. Technically this is equivalent to empirical risk minimization with a regularized error functional within the Frequentist framework.

4. In *Maximum Likelihood Approximation* simply the likelihood of the training data is maximized. This corresponds to a Maximum A Posteriori Approximation with a uniform prior. Technically this is equivalent to empirical risk minimization without regularization terms, being the typical error minimization approach for parametric models.

5. There are situations where even calculating the exact likelihood is too costly. For example, calculating the ϕ–dependent normalization factor $Z(Y|x,\phi)$ of the likelihood can be difficult, or the likelihood may be defined as integral of a nonlinear stochastic differential equation. In such cases the likelihood has to be approximated and the states ϕ (or the parameters ξ) have to be determined by *maximizing an approximate likelihood.* This is sometimes also called *Pseudo Maximum Likelihood Approximation* or *Quasi Maximum Likelihood Approximation* [333]. If prior or regularization terms are included one speaks of a penalized Quasi (or Pseudo) Maximum Likelihood Approximation. (For spatial processes see [21, 42, 43], for a Gaussian approximation in time series analysis [179, 180, 200, 518], for mixture of Gaussians [258], for general nonlinear stochastic differential equations [18, 46, 484], and for a Monte Carlo evaluation of the likelihood see [114, 452].) Another example of a likelihood which can only be calculated in some approximation

[9]The MAP (method of steepest descents, Laplace's approximation, or saddle point approximation) to the predictive density is also sometimes called *penalized Maximum Likelihood Approximation* if the log–prior is merely seen as penalty term.

is provided by Bayesian inverse many–body quantum theory discussed in Section 7.3. For many–body quantum systems possible approximation schemes are provided by mean field theories like the Hartree–Fock approximation.

2.3 *A priori* information

2.3.1 Controlled, measured, and structural priors

The essential role of *a priori* information for learning is to provide the necessary dependencies between test and training data. It is thus fundamental for any nonparametric learning approach to measure or control the dependencies of the test data of interest on the available training data. In the following we discuss how the necessary *a priori* information can be interpreted, obtained, and represented in a statistical model.

Each likelihood $p(y|x,\phi)$, describing a measurement of ϕ under condition x, always contains aspects of measurement and aspects of control. The independent variable x represents an explicit (stochastic) control of y. This dependent variable y contains at the same time information on ϕ, i.e., it measures ϕ. Similarly, (prior) information on ϕ can be interpreted as measurement or control. One possibility is to interpret *a priori* information $D_0 = (x,y)$, like standard training data, as result of a measurement process described by a likelihood $p(y|x,\phi)$. The other possibility is to interpret *a priori* information as a collection of variables which control (stochastically) the state ϕ according to some (prior) process $p(\phi|D_0)$. That part of *a priori* information which is included explicitly in a statistical model can thus appear as

1. (probabilistic) control, or as
2. measurement.

That part of *a priori* information which is not modeled explicitly as measurement or control, but is implicit in the model structure, will be called

3. structural prior.

An example for (an interpretation as) structural priors are models which assume certain factorization properties. Hence *a priori* information can appear in a statistical model in three different forms.

Let us first consider the case of a local bound $y(x) \leq B$. To verify this bound empirically one would have to measure a given state ϕ completely to ensure that no $p(y > B|x,\phi) \neq 0$ appears. A second possibility is to use a measuring device which is only able to produce outcome $y(x) \leq B$. This can be achieved by explicit selection, cutting or squashing. Finally, such a bound arises, for example, in models which assume that ϕ is sampled from a prior

$p(\phi|D_0)$ having zero density for states ϕ which correspond to likelihoods with $p(y > B|x, \phi) \neq 0$. Hence in the first variant we measure ϕ, in the second we explicitly enforce the restriction on y, and in the third we assume a model with a specific sampling process for ϕ. Such model assumptions have of course also to be controlled or measured, but in the third variant this is not formulated explicitly in the model. (The difficulties of controlling $p(\phi|D_0)$, being a more complicated object than ϕ itself, or of 'measuring' a complete function $\phi(x, y)$ are discussed in Appendix A.)

As a second example consider some nonlocal *a priori* information, e.g., on the smoothness of the likelihood. Again, we can either sample ϕ (from a Gaussian process with smoothness related covariance, for example), perform a noisy (nonlocal) measurement of 'smoothness', or use a measurement device which is only able to produce smooth likelihoods, as in the case of input and output noise discussed below.

Let us now clarify the concepts of control and measurement. By a (pure) stochastic control of y we mean choosing x and sampling y according to a known $p(y|x)$. By measuring an unknown ϕ we mean sampling data y according to $p(y|\phi)$. Thus a likelihood $p(y|x, \phi)$ describes aspects of measurement and aspects of control: The outcome y is stochastically controlled by the known input x and unknown state ϕ, so reading off y yields information on that unknown ϕ.

Controlled priors: As the examples have shown, a prior $p(\phi|D_0)$ can be implemented by sampling ϕ according to $p(\phi|D_0)$ controlled by 'prior parameters' D_0. Such a *controlled (sampling) prior* is shown in Fig. 2.3a. Dependencies between the components of ϕ [e.g., the $\phi(x, y)$] are here generated by the sampling process $p(\phi|D_0) = p(\phi_1|D_0)p(\phi_2|\phi_1, D_0)\, p(\phi_3|\phi_1, \phi_2 D_0)\cdots$. (For Gaussian processes (2.106) one can sample independent variables and perform a linear transformation to obtain the desired covariance structure.)

Measured priors: On the other hand measured data can be incorporated in the prior using Bayes' theorem. For example, (standard local) training data obtained in the past may already have been included in the prior. To obtain in a nonparametric setting information on the dependencies between training data and test data requires generalized (nonlocal) measurements. As mentioned in Section 2.1.7, a Gaussian process prior (2.106), for instance, can also be interpreted as likelihood of the measurement of a function $t(x, y)$ (or $t(x)$ for regression problems) under Gaussian noise with real–symmetric, positive definite covariance $\mathbf{K}^{-1}$ as in Eq. (2.114)

$$p(\phi|D_0) = \frac{e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t\rangle}}{\int d\phi'\, e^{-\frac{1}{2}\langle \phi' - t|\mathbf{K}|\phi' - t\rangle}} = \frac{e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t\rangle}}{\int dt'\, e^{-\frac{1}{2}\langle \phi - t'|\mathbf{K}|\phi - t'\rangle}} = p(t|\phi, \mathbf{K}). \tag{2.180}$$

This is possible, because the Gaussian integral $\int dt'\, e^{-\frac{1}{2}\langle \phi - t'|\mathbf{K}|\phi - t'\rangle}$ does not

depend on ϕ. We call a prior interpreted in analogy to $p(y|x,\phi)$ as a (generalized) likelihood $p(t|\mathbf{K},\phi)$ a *measured prior* (see Fig. 2.21a). A positive definite inverse covariance can be decomposed in square roots, $\mathbf{K} = \mathbf{W}^T\mathbf{W}$ (the superscript ^{T} denoting the transpose), hence Eq. (2.180) is also the likelihood for measuring a complete 'filtered' function $\mathbf{W}\phi$ under Gaussian noise. (For positive semi–definite $\mathbf{K}$ one has to take care of zero modes. We understand a positive (semi–)definite operator to be symmetric as well.) The function t in (2.180) is thus a generalization of local training data as they appear in Gaussian regression,

$$p(y|x,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\phi(x))^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\langle y-\phi|\mathbf{K}_x|y-\phi\rangle}, \qquad (2.181)$$

corresponding to an inverse covariance $\mathbf{K}_x(x,x') = \sigma^{-2}\delta(x-x')$. We see that in this Gaussian case choosing to measure y at value x can be interpreted as selecting a specific covariance for the measurement. Note that, in contrast to local training data, generalized data must relate training and test data to be useful for learning. For example, such generalized data must depend on all x which can appear as test data.

It is always possible to interpret a given prior as controlled or as measured prior. In particular, we can separate the available *a priori* information D_0 into a controlled part D_C and a measured part D_M and write

$$p(\phi|D_0) = p(\phi|D_M, D_C) = \frac{p(D_M|\phi, D_C)p(\phi|D_C)}{p(D_M|D_C)}. \qquad (2.182)$$

In case of a uniform $p(\phi|D_C)$ we speak of a completely measured prior.

Structural priors: Besides the possibilities to model *a priori* information explicitly as measurement or control, the relevant dependencies can also just be implemented implicitly in the structure of the model. Choosing this implementation of *a priori* information, dependencies arise because according to the given model the same components of ϕ contribute to the likelihood values $p(y|x,\phi)$ for more than one (x,y) pair. Such a *structural prior* is shown in Figs. 2.3b. and 2.21b. For example in parametric models, where ϕ is represented by a few parameters ξ, those determine all values of $p(y|x,\xi)$. Also in nonparametric models a specific likelihood value $p(y|x,\phi)$ can depend on the whole function ϕ, for example, through the normalization constraint (see Section 3). Similarly, in models with input or output noise, several values $\phi(x,y)$ influence a specific likelihood value of the noisy measurement device. While for a measured prior the dependency information results from a generalized *a priori* measurement, for a structural prior dependencies are implemented by modelling the training data itself as generalized (nonlocal) measurements with respect to the hidden model variables ϕ. In Section 2.3.2 and Appendix A it is discussed how dependency structures can be induced through the measurement process.

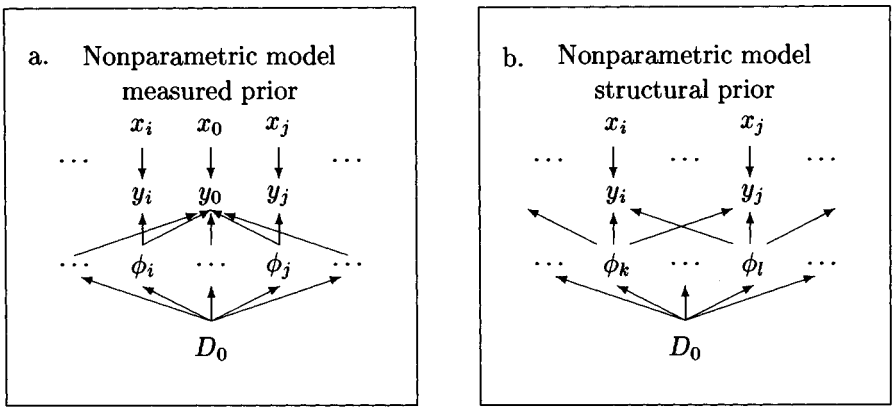


Figure 2.21: Measured and structural dependencies. a. Measured prior: Here information on dependencies between different ϕ_i is obtained from generalized (nonlocal) measurements. For example, x_0 can represent a certain smoothness measure defined by an inverse covariance $\mathbf{K}$, and $y_0 = t(x)$ can be a function of the corresponding measured local smoothness values. b: Same as Fig. 2.3b for parametric approaches. In contrast to a finite parameter vector ξ the letter ϕ is used to denote a function $\phi(x, y)$ or $\phi(x)$. A controlled (sampling) prior is shown in Fig. 2.3a.

One can switch between the three interpretations by either applying Bayes' theorem or by introducing additional hidden variables to implement specific dependencies as model structure.

2.3.2 Noise induced priors

As examples of specific structurally implemented dependencies let us discuss the three cases of input noise (2.19), output noise (2.20), and output averaging with multiple inputs.

Input noise: Consider a realistic measurement device which is intended to measure at position x but is not able to measure always exactly at this position. Instead it picks an $\tilde{x}$, for example, in the neighborhood of x, with some density $p(\tilde{x}|x)$. This situation can be described by a model with input noise. (See also [291] and [53, 286, 514] for a related error minimization approach.) Introducing two different descriptions ('parameterizations') of the same state of Nature, ϕ for the effective likelihood $p(y|x, \phi)$ for the visible x, y, and $\tilde{\phi}$ for the hidden likelihood $p(y|\tilde{x}, \phi)$, model (2.19) becomes

$$p(y|x, \phi) = \int d\tilde{x}\, p(y|\tilde{x}, \tilde{\phi})\, p(\tilde{x}|x). \tag{2.183}$$

As $p(y|x,\phi)$ is determined by $p(y|\tilde{x},\tilde{\phi})$ the description ϕ is a deterministic function of $\tilde{\phi}$ for given $p(\tilde{x}|x)$, so that $p(y|x,\phi) = p(y|x,\phi(\tilde{\phi})) = p(y|x,\tilde{\phi})$. Clearly, the explicit functional dependence of $p(y|x,\phi)$ on ϕ is different from that on $\tilde{\phi}$. Different $\tilde{\phi}$ can lead to the same ϕ. However, ϕ must provide a complete description of the likelihood. That means if two different $\tilde{\phi}$ and $\tilde{\phi}'$ lead to the same state, i.e., $\phi(\tilde{\phi}) = \phi(\tilde{\phi}') = \phi'$, then $p(y|x,\phi) = p(y|x,\phi')$ must hold. This is always fulfilled for nonparametric density estimation if $\phi(x,y) = p(y|x,\phi)$, but not necessarily for regression or parametric approaches. To be specific we discuss the cases of nonparametric regression and general density estimation and show that such input noise can induce dependencies between effective likelihood values $p(y|x,\phi)$ which are not present between hidden likelihood values $p(y|\tilde{x},\tilde{\phi})$.

In regression the form of the likelihood for given x is fixed except for its mean. This leaves a regression function $\phi(x) = \int dy\, y\, p(y|x,\phi)$ to be reconstructed. Defining analogous to $\phi(x)$ the hidden regression function $\tilde{\phi}(\tilde{x})$, the two regression functions

$$\phi(x) = \langle y \rangle_{y|x,\phi} = \int dy\, y\, p(y|x,\phi), \tag{2.184}$$

$$\tilde{\phi}(\tilde{x}) = \langle y \rangle_{y|\tilde{x},\tilde{\phi}} = \int dy\, y\, p(y|\tilde{x},\tilde{\phi}), \tag{2.185}$$

are related according to Eq. (2.183) as

$$\phi(x) = \int d\tilde{x}\, \mathbf{B}_X(x,\tilde{x})\tilde{\phi}(\tilde{x}), \tag{2.186}$$

where

$$\mathbf{B}_X(x,\tilde{x}) = p(\tilde{x}|x). \tag{2.187}$$

Even with the hidden regression function $\tilde{\phi}(\tilde{x})$ providing a complete description of $p(y|\tilde{x},\tilde{\phi})$, the visible regression function $\phi(x)$ does not necessarily provide a complete description of $p(y|x,\phi)$. Thus there can exist visible likelihoods $p(y|x,\phi)$ which are different but have the same regression function $\phi(x)$. Indeed, choosing for example $p(\tilde{x}|x) = 0.5[\delta(\tilde{x}-x) + \delta(\tilde{x}+x)]$ all antisymmetric $\tilde{\phi}(x) = -\tilde{\phi}(-x)$ lead to likelihoods $p(y|x,\phi)$ with $\phi(x) = 0$. The regression function $\phi(x)$ is a sufficient description of the visible likelihood if $\mathbf{B}_X$ has no zero modes so that $\tilde{\phi} \neq \tilde{\phi}' \Rightarrow \phi \neq \phi'$.

Consider now a factorial prior over $\tilde{x}$ for the hidden regression function $\tilde{\phi}$ as in Eq. (2.80), $p(\tilde{\phi}) = \prod_{\tilde{x}} p(\tilde{\phi}(\tilde{x}))$, so the function values $\tilde{\phi}(\tilde{x})$ for different $\tilde{x}$ are independent random variables. Furthermore, assume $\mathbf{B}_X$ is of such a form that a large number of values $\tilde{\phi}(\tilde{x})$ contribute to each value $\phi(x)$ and the $p(\tilde{\phi}(\tilde{x}))$ have existing variance so that the (general) central limit theorem applies. Then $p(\phi)$ is approximately Gaussian [62, 438]. Defining the mean function $t(x) = \int d\phi\, \phi(x) p(\phi)$ of the prior $p(\phi)$ and analogously $\tilde{t}$ of $p(\tilde{\phi})$, we

have

$$t(x) = \int d\phi\, d\tilde{\phi}\, \phi(x) p(\phi|\tilde{\phi}) p(\tilde{\phi}) = \int d\phi\, d\tilde{\phi}\, \phi(x) \delta(\phi - \mathbf{B}_X \tilde{\phi}) p(\tilde{\phi})$$
$$= \int d\tilde{x}\, \mathbf{B}_X(x,\tilde{x}) \int d\tilde{\phi}\, \tilde{\phi}(\tilde{x}) p(\tilde{\phi}) = \int d\tilde{x}\, \mathbf{B}_X(x,\tilde{x}) \tilde{t}(\tilde{x}), \quad (2.188)$$

i.e., $t = \mathbf{B}_X \tilde{t}$. Similarly, we get for the covariances

$$\mathbf{C}(x,x') = \int d\phi\, \phi(x)\phi(x') p(\phi) - t(x)t(x'), \quad (2.189)$$

$\tilde{\mathbf{C}}$ defined analogously,

$$\mathbf{C} = \mathbf{B}_X \tilde{\mathbf{C}} \mathbf{B}_X^T, \quad (2.190)$$

the superscript ^{T} denoting the transpose. We conclude that input noise can generate a Gaussian prior on $\phi(x)$ even if the prior for $\tilde{\phi}$ is factorial, the $\phi(x)$ are correlated while the $\tilde{\phi}(\tilde{x})$ are not.

Similarly one can proceed for nonparametric density estimation. Choosing, for example, as $\phi(x,y)$ the likelihood values themselves

$$\phi(x,y) = p(y|x,\phi), \quad \tilde{\phi}(\tilde{x},y) = p(y|\tilde{x},\tilde{\phi}), \quad (2.191)$$

Eq. (2.183) reads

$$\phi(x,y) = \int d\tilde{x}\, \mathbf{B}_X(x,\tilde{x}) \tilde{\phi}(\tilde{x},y), \quad (2.192)$$

which can be written $\phi(\cdot,y) = \mathbf{B}_X \tilde{\phi}(\cdot,y)$ where the integration is understood to be restricted to $\tilde{x}$. Thus a factorial prior over x leads, using the same arguments as above, for an appropriate $\mathbf{B}_X$ and finite covariance of $\tilde{\phi}$, to a prior for $\phi(x,y)$ which is approximately Gaussian over x (but not over y) according to the central limit theorem. For prior mean $t(x,y) = \int d\phi\, \phi(x,y) p(\phi)$ and prior covariance $\mathbf{C}(x,y;x',y') = \int d\phi\, \phi(x,y)\phi(x',y') p(\phi) - t(x,y)t(x',y')$, $\tilde{t}$, $\tilde{\mathbf{C}}$ defined analogously, we find

$$t(\cdot,y) = \mathbf{B}_X \tilde{t}(\cdot,y) \quad (2.193)$$

and

$$\mathbf{C}(\cdot,y;\cdot,y') = \mathbf{B}_X \tilde{\mathbf{C}}(\cdot,y;\cdot,y') \mathbf{B}_X^T. \quad (2.194)$$

Thus also in a nonparametric density estimation framework input noise leads to dependencies of likelihood values with different x.

Output noise: While input noise can relate x–values of likelihoods, output noise (2.20) can relate y–values. For nonparametric density estimation an output noise model

$$p(y|x,\phi) = \int d\tilde{y}\, p(y|\tilde{y})\, p(\tilde{y}|x,\tilde{\phi}) \quad (2.195)$$

with $\phi(x,y) = p(y|x,\phi)$ and $\tilde{\phi}(x,\tilde{y}) = p(\tilde{y}|x,\tilde{\phi})$ reads $\phi(x,\cdot) = \mathbf{B}_Y\tilde{\phi}(x,\cdot)$ or shorter, $\phi(x) = \mathbf{B}_Y\tilde{\phi}(x)$, where, in contrast to Eq. (2.187), $\mathbf{B}_Y(y,\tilde{y}) = p(y|\tilde{y})$. To proceed we assume that the observed variable y can take many different values, hence dependencies due to the combined non–negativity and normalization constraints can be neglected. Then choosing $p(\tilde{\phi})$ as an (approximately) factorial prior, the hidden likelihoods $\tilde{\phi}(x,\tilde{y})$ represent (approximately) independent variables. Like for input noise, the prior of the likelihoods $p(\phi)$ becomes for appropriate $\mathbf{B}_Y$ approximately Gaussian over y if many values $\tilde{\phi}(x,\tilde{y})$ contribute to one $\phi(x,y)$, with finite covariance of the components, so that the general central limit theorem applies. Prior mean and covariance are obtained as $t(x,\cdot) = \mathbf{B}_Y\tilde{t}(x,\cdot)$ and $\mathbf{C}(x,\cdot;x',\cdot) = \mathbf{B}_Y\tilde{\mathbf{C}}(x,\cdot;x',\cdot)\mathbf{B}_Y^T$.

Priors for likelihoods which are approximately Gaussian with respect to both x and y can be obtained by combining input and output noise.

Output averaging with multiple inputs: Like noise processes, averaging processes can also induce dependencies and at the same time effectively reduce the variance of the observed values. Consider, for example, a measurement device which obtains first a vector $\vec{\tilde{y}}$ of (hidden) measurements $\tilde{y}_j$ at positions $\tilde{x}_j$ (the latter combined into a vector $\vec{\tilde{x}}$, possibly depending on x) and then returns as observed value y a (probabilistic) average of the components $\tilde{y}_j$ described by $p(y|x,\vec{\tilde{y}})$ (see Fig. 2.22),

$$p(y|x,\phi) = \int d\vec{\tilde{y}}\, d\vec{\tilde{x}}\, p(y|x,\vec{\tilde{y}}) \left(\prod_j p(\tilde{y}_j|\tilde{x}_j,\tilde{\phi}) \right) p(\vec{\tilde{x}}|x). \tag{2.196}$$

In case the central limit theorem applies the averaging will produce Gaussian likelihoods, while including values from different $\tilde{x}$ induces dependencies of the effective likelihoods $p(y|x,\phi)$ over x.

Now consider as special case a measurement device which for each observation at some x samples a (hidden) value $\tilde{y}(\tilde{x})$ for each possible value of $\tilde{x}$, so $\vec{\tilde{y}}$ becomes a function $\tilde{y}(\tilde{x})$ and returns as observed y at x the value of the linear combination

$$y(x) = \int d\tilde{x}\, \mathbf{L}_X(x,\tilde{x})\, \tilde{y}(\tilde{x}). \tag{2.197}$$

Here we have in mind for $\mathbf{L}_X$ some averaging process, for example, with $\mathbf{L}_X \geq 0$ and $\int d\tilde{x}\, \mathbf{L}_X(x,\tilde{x}) = 1$. Introducing analogous to $\vec{\tilde{y}} = \tilde{y}(\cdot)$ the function $\vec{y} = y(\cdot)$ — from which actually only the value at x is observed — we can write $\vec{y} = \mathbf{L}_X\vec{\tilde{y}}$. Using Eq. (2.197) we have to insert in Eq. (2.196)

$$p(y|x,\vec{\tilde{y}}) = \delta\Big(y - \int d\tilde{x}\, \mathbf{L}_X(x,\tilde{x})\tilde{y}(\tilde{x})\Big), \tag{2.198}$$

$$p(\vec{\tilde{x}}|x) = \prod_{\tilde{x}'} \delta(\tilde{x} - \tilde{x}'), \tag{2.199}$$

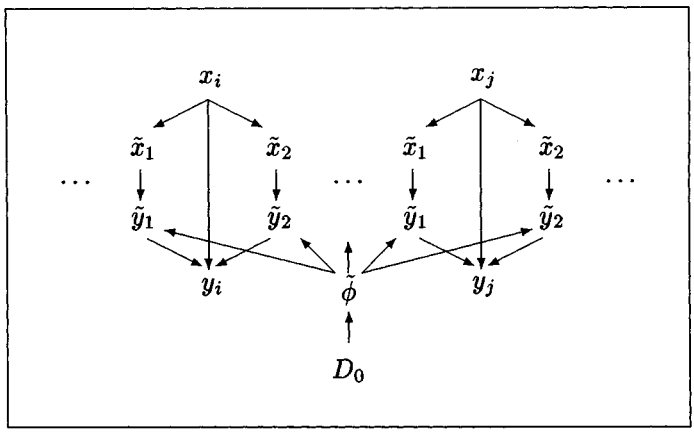


Figure 2.22: Output averaging with multiple inputs. In this example $\vec{\tilde{x}} = (\tilde{x}_1, \tilde{x}_2)$ and $\vec{\tilde{y}} = (\tilde{y}_1, \tilde{y}_2)$ are two-dimensional.

where the product is over all values $\tilde{x}'$ the random variable $\tilde{X}$ can take, so $p(\vec{\tilde{x}}|x)$ is independent of x. We obtain

$$p(y|x,\phi) = \int d\vec{\tilde{y}}\, \delta\Big(y - \int d\tilde{x}\, \mathbf{L}_X(x,\tilde{x})\tilde{y}(\tilde{x})\Big) \left(\prod_{\tilde{x}} p(\tilde{y}(\tilde{x})|\tilde{\phi})\right), \qquad (2.200)$$

where $p(\tilde{y}(\tilde{x})|\tilde{\phi}) = p(\tilde{y}|\tilde{x},\tilde{\phi})$. In terms of the functions $\vec{y} = y(\cdot)$, $\vec{\tilde{y}} = \tilde{y}(\cdot)$, Eq. (2.200) reads $p(\vec{y}|\phi) = \int d\vec{\tilde{y}}\, \delta(\vec{y} - \mathbf{L}_X\vec{\tilde{y}}) p(\vec{\tilde{y}}|\tilde{\phi})$. (Averaged over many hidden measurements for the same $\tilde{x}$, the function $\tilde{y}(\tilde{x})$ becomes asymptotically the regression function $\tilde{\phi}(\tilde{x})$. Such an averaging of an underlying regression function is shown in Fig. 2.23.)

Again, the central limit theorem applies to y of Eq. (2.197) if the variances of $p(\tilde{y}|\tilde{x},\tilde{\phi})$ are finite and $\mathbf{L}_X$ is an averaging operator so that many components contribute to y. Mean $t_y(x) = \langle y(x)\rangle_{\vec{y}|\phi} = \int d\vec{y}\, y(x) p(\vec{y}|\phi)$ and covariance $\mathbf{C}_y(x,x') = \langle y(x)y(x')\rangle_{\vec{y}|\phi} - \langle y(x)\rangle_{\vec{y}|\phi}\langle y(x')\rangle_{\vec{y}|\phi} = \int d\vec{y}\, y(x)y(x') p(\vec{y}|\phi) - t_y(x)t_y(x')$ of the Gaussian $p(\vec{y}|\phi)$ are related to the analogous $\tilde{t}_{\tilde{y}}(\tilde{x})$, $\tilde{\mathbf{C}}_{\tilde{y}}(\tilde{x},\tilde{x}')$ as follows

$$t_y = \mathbf{L}_X \tilde{t}_{\tilde{y}}, \quad \mathbf{C}_y = \mathbf{L}_X \tilde{\mathbf{C}}_{\tilde{y}} \mathbf{L}_X^T. \qquad (2.201)$$

Here the covariance $\tilde{\mathbf{C}}_{\tilde{y}}$ is diagonal for $p(\vec{\tilde{y}}|\tilde{\phi}) = \prod_{\tilde{x}} p(\tilde{y}(\tilde{x})|\tilde{\phi})$ as in Eq. (2.200).

Hence let us consider the asymptotic case of a Gaussian likelihood. Such likelihoods can be described completely by giving regression function and covariance. Let us furthermore assume that the regression function $\phi(x) = t_y(x)$ describes the likelihood completely. Then Eq. (2.201) reads $\phi = \mathbf{L}_X \tilde{t}_{\tilde{y}}$. Also, if the hidden likelihood is determined by its regression function $\tilde{\phi}(\tilde{x})$

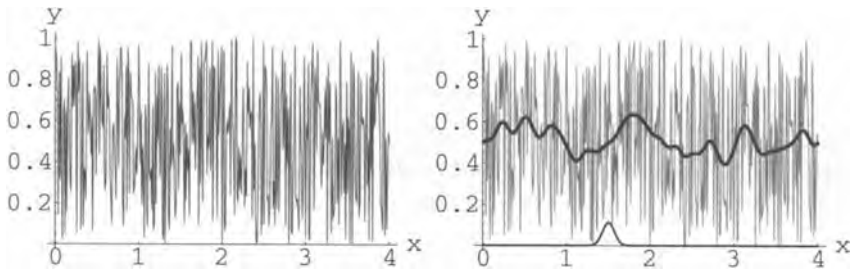


Figure 2.23: Noise and learning. The left hand side shows a bounded random function which does not allow generalization from training to non–training data. However, using a realistic measurement device with input noise or averaging, the function becomes learnable. The right hand side shows the corresponding regression function as seen by a measurement device which implements a binomial filtering (shown at the bottom) of the original regression function. The fact that such a realistic measurement device can not resolve neighboring x–values perfectly, results in a smooth effective regression function which induces dependencies between different x–values and thus allows generalization from training to non–training data. While in classical inverse problems one is aiming at inverting a filter ('instrument or apparatus function') to reconstruct underlying original data, here the role of the filter is to allow learning from filtered (derived, correlated) data which would not be possible from the original underlying random data.

$= \tilde{t}_{\tilde{y}}(\tilde{x})$ this becomes $\phi = \mathbf{L}_X \tilde{\phi}$. Similar to input noise, this requires that equal regression functions $\phi = \phi'$ correspond to equal likelihoods $p(y|x,\phi) = p(y|x,\phi')$. In particular, to each regression function has to exist a unique and known covariance. This is the case, if the hidden likelihoods $p(\tilde{y}|\tilde{x},\tilde{\phi})$ are already completely determined by their regression function $\tilde{\phi}(\tilde{x})$, so that no hidden likelihoods appear with equal $\tilde{\phi}(\tilde{x})$, and if $\mathbf{L}_X$ is invertible, so that $\tilde{\phi} \neq \tilde{\phi}' \Rightarrow \phi(\tilde{\phi}) \neq \phi(\tilde{\phi}')$ and thus also no likelihoods with equal $\phi(x)$ occur.

As discussed for input noise, the linear relationship $\phi = \mathbf{L}_X \tilde{\phi}$ leads, for a factorial prior over $\tilde{\phi}(\tilde{x})$ with existing covariance and appropriate averaging $\mathbf{L}_X$, according to the central limit theorem to a Gaussian prior for the (visible) regression function $\phi(x)$. Again visible and hidden prior mean and prior covariance are related as $t = \mathbf{L}_X \tilde{t}$ and $\mathbf{C} = \mathbf{L}_X \tilde{\mathbf{C}} \mathbf{L}_X^T$.

Summarizing, Gaussian regression problems with Gaussian prior on the regression function can result when a measurement device is used which effectively performs an (x–dependent) average over some hidden underlying conditions $\tilde{x}$.

We conclude that imperfect, noisy or averaging measurement devices introduce dependencies between likelihoods. Such dependencies, however, enable learning. *From this point of view, the ability to learn arises from the inability to measure exactly.* In other words, realistic measurement devices being not able to distinguish always perfectly different values of their input and output variables, induce dependencies between likelihood values which enable generalization from training to test data.

Noise and symmetries: It is also instructive to formulate noise processes in terms of a family of transformations S_X or S_Y acting on x or $\tilde{y}$, respectively. For $\tilde{x}$ given by a function $S_X(x)$ of x or, equivalently, in operator notation with $\mathbf{S}_X$ acting on x,

$$\tilde{x} = S_X(x) = \mathbf{S}_X x, \tag{2.202}$$

the integral over $\tilde{x}$ can be replaced by an integral over transformations S_X or operators $\mathbf{S}_X$. The linear operator $\mathbf{S}_X$ acts on x as defined by the function S_X when x is considered as element in a function space with coordinate values $x(x') = \langle x' | x \rangle = \delta(x - x')$. For symmetry transformations S_X is one–to–one, the corresponding operator $\mathbf{S}_X = \int dx \, |S_X(x)\rangle \langle x|$, being for discrete spaces a permutation operator, has inverse $\mathbf{S}_X^{-1} = \mathbf{S}_X^T$ and thus $|\det \mathbf{S}_X| = 1$. Its matrix elements are $\langle x' | \mathbf{S}_X | x \rangle = \int dx'' \, \mathbf{S}_X(x', x'') \delta(x'' - x) = \mathbf{S}_X(x', x) = \delta(x' - S_X(x)) = |\det \mathbf{S}_X^{-1}| \delta(\mathbf{S}_X^{-1} x' - x) = \delta(\mathbf{S}_X^{-1} x' - x)$. Analogously, we can write for output noise $y = S_Y(\tilde{y}) = \mathbf{S}_Y \tilde{y}$.

Introducing $p(\mathbf{S}_X|x)$ as the probability density for a symmetry transformation $\mathbf{S}_X$ to act during the measurement process on x, input noise can be formulated in terms of transformations $\mathbf{S}_X$ (and similarly output noise in terms of $\mathbf{S}_Y$). Indeed, because a scalar function fulfills $f(x) = f'(x') = \mathbf{S}_X f(\mathbf{S}_X x)$, and thus $f(\mathbf{S}_X x) = \mathbf{S}_X^{-1} f(x)$, we can let the operator $\mathbf{S}_X^{-1}$ act on

the likelihood instead of $\mathbf{S}_X$ on x and obtain

$$\begin{aligned} p(y|\tilde{x},\tilde{\phi}) &= p(y|S_X(x),\tilde{\phi}) = p(y|\mathbf{S}_\mathbf{X}x,\tilde{\phi}) = (\mathbf{S}_X^{-1}p)(y|x,\tilde{\phi}) \qquad (2.203)\\ &= \int dx'\, \mathbf{S}_X^{-1}(x,x')p(y|x',\tilde{\phi}) = \int dx'\, \mathbf{S}_X^{-1}(S_X^{-1}(\tilde{x}),x')p(y|x',\tilde{\phi}). \end{aligned}$$

Thus with deterministic $p(\tilde{x}|S_X(x)) = p(\tilde{x}|\mathbf{S}_X,x) = \delta(\tilde{x}-S_X(x)) = \delta(\tilde{x}-\mathbf{S}_X x)$ one gets for Eq. (2.183)

$$\begin{aligned} p(y|x,\phi) &= \int d\tilde{x}\, p(y|\tilde{x},\tilde{\phi})p(\tilde{x}|x) = \int d\mathbf{S}_X\, d\tilde{x}\, \mathbf{S}_X^{-1}(x,\tilde{x})p(y|\tilde{x},\tilde{\phi})p(\mathbf{S}_X|x)\\ &= \int d\mathbf{S}_X\, p(y|\mathbf{S}_X,x,\tilde{\phi})p(\mathbf{S}_X|x). \qquad (2.204) \end{aligned}$$

For example, in case of a Lie group $\mathbf{S}_X(\alpha)$ the integral over $\mathbf{S}_X$ stands for an integral over parameters α. For nonparametric density estimation with $\phi(x,y) = p(y|x,\phi)$ Eq. (2.204) yields again Eq. (2.192),

$$\phi = \mathbf{B}_X\tilde{\phi}, \qquad (2.205)$$

where, expressed now in terms of $\mathbf{S}_X$,

$$\mathbf{B}_X(x,\tilde{x}) = \int d\mathbf{S}_X\, p(\mathbf{S}_X|x)\mathbf{S}_X^{-1}(x,\tilde{x}) = \int d\mathbf{S}_X\, p(\mathbf{S}_X|x)\delta(\tilde{x}-S_X(x)) = p(\tilde{x}|x). \qquad (2.206)$$

Similarly, one finds for output noise

$$p(y|x,\phi) = \int d\tilde{y}\, p(y|\tilde{y})\, p(\tilde{y}|x,\tilde{\phi}) = \int d\mathbf{S}_Y p(\mathbf{S}_Y|\mathbf{S}_Y^{-1}y,x)\, p(\mathbf{S}_Y^{-1}y|x,\tilde{\phi}), \qquad (2.207)$$

so that

$$\phi = \mathbf{B}_Y\tilde{\phi} \qquad (2.208)$$

with

$$\mathbf{B}_Y(y,\tilde{y}) = \int d\mathbf{S}_Y\, p(\mathbf{S}_Y|\tilde{y},x)\mathbf{S}_Y(y,\tilde{y}) = \int d\mathbf{S}_Y\, p(\mathbf{S}_Y|\tilde{y})\delta(y-S_Y(\tilde{y})) = p(y|\tilde{y}). \qquad (2.209)$$

Noise processes can produce symmetric likelihoods. For example $\mathbf{B}_X(x,x) = \mathbf{B}_X(x,-x) = 0.5$ results in a likelihood invariant under reflection $S(x) = -x$. In general, consider a set of transformations $\mathbf{S}$ (which can stand for $\mathbf{S}_X$ or $\mathbf{S}_Y$) which form a symmetry group. A (visible) likelihood $p(y|x,\phi)$ which is symmetric under the considered symmetry group can be obtained by choosing for $\mathbf{B}$ (which can stand for $\mathbf{B}_X$ or $\mathbf{B}_Y$) as $p(\mathbf{S})$ the invariant measure of the

group.[10] Indeed, in that case

$$\mathbf{SB} = \mathbf{S}\int d\mathbf{S}'\, p(\mathbf{S}')\mathbf{S}' = \int d\mathbf{S}'\, p(\mathbf{S}')\mathbf{S}\mathbf{S}' = \int d\mathbf{S}'\, p(\mathbf{S}')\mathbf{S}' = \mathbf{B} \tag{2.210}$$

which means $\mathbf{SB}f = \mathbf{B}f$, hence a function of the form $\mathbf{B}f$ is invariant under all $\mathbf{S}$. The operator $\mathbf{B} = \int d\mathbf{S}\, p(\mathbf{S})\, \mathbf{S} = \int d\mathbf{S}\, p(\mathbf{S})\, \mathbf{S}^{-1} = \int d\mathbf{S}\, p(\mathbf{S})\, \mathbf{S}^T = \mathbf{B}^T$ is symmetric, as follows for an invariant measure from $\mathbf{S}^{-1} = \mathbf{S}^T$. Because of $\mathbf{B}^2 = \int d\mathbf{S}\, d\mathbf{S}'\, p(\mathbf{S})p(\mathbf{S}')\mathbf{S}\mathbf{S}' = \int d\mathbf{S}\, d\mathbf{S}'\, p(\mathbf{S})p(\mathbf{S}')\mathbf{S} = \int d\mathbf{S}\, p(\mathbf{S})\mathbf{S} = \mathbf{B}$, the matrix $\mathbf{B}$ is for an invariant measure a projector to the space of symmetric functions and in general not invertible.

As another example let us consider an invertible $\mathbf{B}$. We have just seen that then the likelihood can only be approximately symmetric under the symmetry group over which is integrated in $\mathbf{B}$. We are then interested in calculating the inverse $\mathbf{B}^{-1} = \mathbf{W}$ from which we can obtain the inverse prior covariance of the likelihood ϕ as in Eq. (2.201). For example, for a unit covariance $\tilde{\mathbf{C}} = \mathbf{I}$ of the hidden likelihood $\tilde{\phi}$ we have $\mathbf{C} = \mathbf{B}\mathbf{B}^T$ so the inverse covariance of the visible likelihood ϕ is given by $\mathbf{K} = \mathbf{W}^T\mathbf{W}$. We call $\mathbf{W}$ a (prior) filter operator.

Next, let us separate in $\mathbf{B}$ a part proportional to the identity $\mathbf{I}$ from a nondiagonal part proportional to $\mathbf{G}$ representing the noise, $\mathbf{B} = a(\mathbf{I} - b\mathbf{G})$. Then, in case the following von Neumann series converges, the filter operator $\mathbf{W} = \mathbf{B}^{-1}$ is given by

$$\mathbf{W} = \frac{1}{a}(\mathbf{I} - b\mathbf{G})^{-1} = \frac{1}{a}\sum_{n=0}^{\infty}(b\mathbf{G})^n = \frac{1}{a}\left(\mathbf{I} + \sum_{n=1}^{\infty}(b\mathbf{G})^n\right). \tag{2.211}$$

Operators $\mathbf{B}$ which represent $\mathbf{B}_X = p(\tilde{x}|x)$ or $\mathbf{B}_Y = p(y|\tilde{y})$ have to satisfy a normalization condition over rows or columns, respectively. If the corresponding normalization condition is fulfilled for $\mathbf{G}$ and $a = 1/(1-b)$, it is also fulfilled for $\mathbf{B}$.

Similarly, we may ask which $\mathbf{B}$ corresponds to a prior filter of the form $\mathbf{W} = \mathbf{B}^{-1} = a'(\mathbf{I} - b'\mathbf{G}')$. In that case we find analogously for $\mathbf{B}$

$$\mathbf{B} = \frac{1}{a'}\sum_{n=0}^{\infty}(b'\mathbf{G}')^n = \frac{1}{a'}\left(\mathbf{I} + \sum_{n=1}^{\infty}(b'\mathbf{G}')^n\right). \tag{2.212}$$

Again, the normalization conditions for $\mathbf{B}_X = p(\tilde{x}|x)$ or $\mathbf{B}_Y = p(y|\tilde{y})$ can be fulfilled, if they are already fulfilled for $\mathbf{G}'$ by choosing again $a' = 1/(1-b')$.

[10] A measure is invariant if $\int d\mathbf{S}\, p(\mathbf{S})\, f(\mathbf{S}) = \int d\mathbf{S}\, p(\mathbf{S})\, f(\mathbf{S}\mathbf{S}') = \int d\mathbf{S}\, p(\mathbf{S})\, f(\mathbf{S}'\mathbf{S})$. Then $\int d\mathbf{S}\, p(\mathbf{S})\, f(\mathbf{S}) = \int d\mathbf{S}\, p(\mathbf{S})\, f(\mathbf{S}^{-1})$, thus $\mathbf{B}_X$ and $\mathbf{B}_Y$ can be treated on the same footing. A unique invariant measure exists for any compact group. For a finite group the equally weighted sum over all group elements is an invariant measure.

Indeed, if $\sum_i \mathbf{G}'_{ij} = 1, \forall j$, we have, written for discrete indices

$$\begin{aligned} 1 &= \sum_{i=1}^{d} \mathbf{B}_{ij} = \frac{1}{a'} \sum_{i=1}^{d} \sum_{n=0}^{\infty} (b')^n \left(\mathbf{G}'^n\right)_{ij} \\ &= \frac{1}{a'} + \frac{b'}{a'} \sum_{i=1}^{d} \mathbf{G}'_{ij} + \frac{1}{a'} \sum_{n=2}^{\infty} (b')^n \sum_{i=1}^{d} \sum_{j_1=1}^{d} \cdots \sum_{j_{n-1}=1}^{d} \mathbf{G}'_{ij_1} \cdots \mathbf{G}'_{j_{n-1}j} \\ &= \frac{1}{a'} \sum_{n=0}^{\infty} (b')^n = \frac{1}{a'(1-b')}, \end{aligned} \tag{2.213}$$

from which the above relation follows. In the limit $b' \to 1$ this yields $1/a' \to 0$ and thus $\mathbf{B} \to 0$. Nevertheless prior filters of the form $\mathbf{W} = a'(\mathbf{I} - \mathbf{S})$, where $\mathbf{G}' = \mathbf{S}$ and $b' = 1$, are useful. Clearly, such a filter is noninvertible as all symmetric functions f, which fulfill $f = \mathbf{S}f$ for all $\mathbf{S}$ in $\mathbf{B}$, are zero modes of $\mathbf{W}$, i.e., $\mathbf{W}f = 0$. Therefore a Gaussian prior as in Eq. (2.106), with zero or symmetric mean function $t = \mathbf{S}t$ and corresponding inverse covariance $\mathbf{K} = \mathbf{W}^T\mathbf{W}$, is maximal for likelihoods invariant under $\mathbf{S}$. Invertible posterior covariances are typically obtained again when training data terms are included. Prior filters will be used in Section 3.6.1 to implement various approximate symmetries.

Summarizing, we can say that input noise and output noise result in approximate (or exact) invariance or 'generalized smoothness' of the visible likelihood under a group of symmetry transformations $\mathbf{S}$. In this interpretation learning becomes possible if a measurement device is not able to distinguish perfectly between a given input x and a transformed $\mathbf{S}_X x$ or a given outcome y and a transformed $\mathbf{S}_Y y$. Again we may say the possibility to predict is based on dependencies induced by imperfect measurement devices.

Chapter 3

Gaussian prior factors

In this chapter nonparametric density estimation problems will be studied working with Gaussian prior factors. The aim is to show that Gaussian prior factors are not only convenient from a technical point of view, but are also quite flexible and can be adapted in many ways to a specific learning task. Using this flexibility to implement available *a priori* information is usually the most important step in practical applications of nonparametric approaches. The first sections will present several possible variants to use Gaussian prior factors for general nonparametric density estimation problems. After showing how problem specific *a priori* information can be implemented by choosing the mean and covariance of the Gaussian prior factor, the specific likelihood models of nonparametric regression, including cluster regression, and classification are investigated. (The combination of the Gaussian prior models with the likelihood model of inverse quantum theory is discussed in Chapter 7.)

3.1 Gaussian prior factor for log–likelihoods

3.1.1 Lagrange multipliers: Error functional $E(L)$

As a first example, we will consider in this section density estimation problems with a Gaussian prior factor in the log–likelihoods, $\ln p(y|x,\phi) = L(x,y)$, followed by a discussion of Gaussian prior factors in the likelihoods $p(y|x,\phi) = P(x,y)$ in the next section. Having studied those two special cases, the approach is generalized in the following sections to Gaussian prior factors on arbitrary functions $\phi(x,y)$ which determine likelihoods.

We formulate nonparametric density estimation in terms of the log–likelihood, choosing as hidden variables the field $\phi(x,y) = \ln p(y|x,\phi) = \ln p(y|x,L) = L(x,y)$. A Gaussian prior factor in L corresponds to a quadratic energy

(or error) term in L,

$$\frac{1}{2}\langle L|\mathbf{K}|L\rangle = \frac{1}{2}||L||^2_{\mathbf{K}} = \frac{1}{2}\int dx\,dy\,dx'dy' L(x,y)\mathbf{K}(x,y;x',y')L(x',y'), \quad (3.1)$$

with operator $\mathbf{K}$ symmetric and positive semi–definite. For positive (semi–) definite $\mathbf{K}$ the scalar product defines a (semi–)norm by $||L||_{\mathbf{K}} = \sqrt{\langle L|\mathbf{K}|L\rangle}$ and a distance by $||L - L'||_{\mathbf{K}}$. The quadratic error term (3.1) corresponds to a Gaussian factor of the prior density, $p(\phi|\tilde{D}_0) = p(L|\tilde{D}_0)$ for L, which has been called in Section 2.1.7 the specific prior. We will consider the posterior density

$$p(\phi|f) \propto$$
$$e^{\sum_i L(x_i,y_i) - \frac{1}{2}\int dx dy dx' dy' L(x,y)\mathbf{K}(x,y;x',y')L(x',y') + \int dx\,\Lambda_X(x)\left(1 - \int dy\, e^{L(x,y)}\right)}, \quad (3.2)$$

including the Lagrange multiplier function $\Lambda(x)$ to ensure normalization. Notice that both, the likelihood $\sum_i L(x_i, y_i)$ and the complete prior, are usually *not* Gaussian due to the presence of the normalization conditions. (An exception is Gaussian regression, see Section 3.8.) The posterior (3.2) corresponds to an error functional

$$E(L) = \beta E_{\text{comb}} = -\langle L|N\rangle + \frac{1}{2}\langle L|\mathbf{K}|L\rangle + \langle e^L - \delta(y)|\Lambda_X\rangle, \quad (3.3)$$

with *data vector (function)*

$$N(x,y) = \sum_i^n \delta(x - x_i)\delta(y - y_i), \quad (3.4)$$

Lagrange multiplier vector (function) $\Lambda_X(x,y) = \Lambda_X(x)$, log–likelihood vector (function) $L(x,y)$, *likelihood vector (function)* $e^L(x,y) = e^{L(x,y)} = P(x,y) = p(y|x,\phi)$, and $\delta(y)(x,y) = \delta(y)$. According to Eq. (3.4)

$$\frac{N(x,y)}{n} = \frac{1}{n}\sum_i^n \delta(x - x_i)\delta(y - y_i) = P_{\text{emp}}(x,y) \quad (3.5)$$

is an *empirical density function* for the joint probability (density) $p(x,y|\phi)$.

Before proceeding we define some notations. Functions of vectors (functions) and matrices (operators), different from multiplication, will be understood element-wise like for example $(e^L)(x,y) = e^{L(x,y)}$. Only multiplication of matrices (operators) will be interpreted as matrix product. Element-wise multiplication has thus to be written with the help of diagonal matrices, built from vectors (functions), denoted by the corresponding bold letters. For instance,

$$\mathbf{I}(x,y;x',y') = \delta(x - x')\delta(y - y'), \quad (3.6)$$

$$\begin{aligned}
\mathbf{L}(x,y;x',y') &= \delta(x-x')\delta(y-y')L(x,y), && (3.7)\\
\mathbf{P}(x,y;x',y') &= \delta(x-x')\delta(y-y')P(x,y)\\
= \mathbf{e}^{\mathbf{L}}(x,y;x',y') &= \delta(x-x')\delta(y-y')e^{L(x,y)}, && (3.8)\\
\mathbf{N}(x,y;x',y') &= \delta(x-x')\delta(y-y')\,N(x,y), && (3.9)\\
\mathbf{\Lambda}_X(x,y;x',y') &= \delta(x-x')\delta(y-y')\Lambda_X(x). && (3.10)
\end{aligned}$$

Being diagonal all these matrices commute with each other. *Element-wise multiplication* can now be expressed as

$$\begin{aligned}
(\mathbf{KL})(x',y';x,y) &= \int dx''dy''\mathbf{K}(x',y';x'',y'')\mathbf{L}(x'',y'';x,y)\\
&= \mathbf{K}(x',y';x,y)L(x,y), && (3.11)
\end{aligned}$$

which in general is not equal to $L(x',y')\mathbf{K}(x',y';x,y)$. In contrast, the product $\mathbf{K}L$ of matrix $\mathbf{K}$ with vector L, does not depend on x, y anymore

$$(\mathbf{K}L)(x',y') = \int dx\,dy\,\mathbf{K}(x',y';x,y)L(x,y). \tag{3.12}$$

The *tensor product* or outer product,

$$(\mathbf{K}\otimes L)(x'',y'';x,y;x',y') = \mathbf{K}(x'',y'';x',y')L(x,y), \tag{3.13}$$

depends on additional x'', y''. Thus introducing

$$I(x,y) = 1, \tag{3.14}$$

we can write

$$L = \mathbf{L}I, \quad P = \mathbf{P}I, \quad e^L = \mathbf{e}^{\mathbf{L}}I, \quad N = \mathbf{N}I, \quad \Lambda_X = \mathbf{\Lambda}_X I. \tag{3.15}$$

Taking now the variational derivative (Fréchet derivative) with respect to $L(x,y)$ of (3.2), or equivalently of (3.3), using

$$\frac{\delta L(x',y')}{\delta L(x,y)} = \delta(x-x')\delta(y-y'), \tag{3.16}$$

and setting the gradient equal to zero, yields the stationarity equation

$$0 = N - \mathbf{K}L - \mathbf{e}^{\mathbf{L}}\Lambda_X, \tag{3.17}$$

where $\mathbf{e}^{\mathbf{L}}\Lambda_X = \mathbf{\Lambda}_X e^L = \mathbf{P}\Lambda_X$. The Lagrange multiplier function Λ_X is determined by the normalization condition

$$Z_X(x) = \int dy\, e^{L(x,y)} = 1, \quad \forall x \in X, \tag{3.18}$$

which can also be written as

$$Z_X = \mathbf{I}_X P = \mathbf{I}_X e^L = I \quad \text{or} \quad \mathbf{Z}_X = \mathbf{I}, \tag{3.19}$$

in terms of normalization vector, $Z_X(x,y) = Z_X(x)$, normalization matrix, $\mathbf{Z_X}(x,y;x',y') = \delta(x-x')\delta(y-y')\,Z_X(x)$, and identity on X, $\mathbf{I}_X(x,y;x',y') = \delta(x-x')$. Multiplication of a vector with $\mathbf{I}_X$ corresponds to y–integration. Being a non–diagonal matrix, $\mathbf{I}_X$ does not, in general, commute with diagonal matrices like $\mathbf{L}$ or $\mathbf{P}$. Note also that despite $\mathbf{I}_X e^L = \mathbf{I}_X \mathbf{e^L} I = \mathbf{I} I = I$, in general $\mathbf{I}_X \mathbf{P} = \mathbf{I}_X \mathbf{e^L} \neq \mathbf{I} = \mathbf{Z}_X$. According to the fact that $\mathbf{I}_X$ and $\mathbf{\Lambda}_X$ commute,

$$[\mathbf{\Lambda}_X, \mathbf{I}_X] = \mathbf{\Lambda}_X \mathbf{I}_X - \mathbf{I}_X \mathbf{\Lambda}_X = 0, \tag{3.20}$$

$[A,B]$ denoting the commutator, and that the same holds for the diagonal matrices $[\mathbf{\Lambda}_X, \mathbf{e^L}] = [\mathbf{\Lambda}_X, \mathbf{P}] = 0$, it follows from the normalization condition $\mathbf{I}_X P = I$ that

$$\mathbf{I}_X \mathbf{P} \Lambda_X = \mathbf{I}_X \mathbf{\Lambda_X} P = \mathbf{\Lambda_X} \mathbf{I}_X P = \mathbf{\Lambda_X} I = \Lambda_X, \tag{3.21}$$

i.e.,

$$0 = (\mathbf{I} - \mathbf{I}_X \mathbf{e^L}) \Lambda_X = (\mathbf{I} - \mathbf{I}_X \mathbf{P}) \Lambda_X. \tag{3.22}$$

For $\Lambda_X(x) \neq 0$ Eqs. (3.21) and (3.22) are equivalent to the normalization (3.18). If there exist directions at the stationary point L^* in which the normalization of P changes, i.e., the normalization constraint is active, $\Lambda_X(x) \neq 0$ restricts the gradient to the normalized subspace (Kuhn–Tucker conditions [40, 134, 223, 402]). This will clearly be the case for the unrestricted variations of $p(y,x)$ which we are considering here. Combining $\Lambda_X = \mathbf{I}_X \mathbf{P} \Lambda_X$ for $\Lambda_X(x) \neq 0$ with the stationarity Eq. (3.17), the Lagrange multiplier function is obtained

$$\Lambda_X = \mathbf{I}_X (N - \mathbf{K}L) = N_X - (\mathbf{I}_X \mathbf{K} L). \tag{3.23}$$

Here we introduced the vector $N_X = \mathbf{I}_X N$, with components

$$N_X(x,y) = N_X(x) = \sum_i \delta(x - x_i) = n_x, \tag{3.24}$$

giving the number of data available for x. Thus Eq. (3.23) reads in components

$$\Lambda_X(x) = \sum_i \delta(x - x_i) - \int dy''\, dx' dy'\, \mathbf{K}(x,y'';x',y') L(x',y'). \tag{3.25}$$

We remark, that for a Laplacian $\mathbf{K}$ and cyclic or zero boundary conditions, the integral term in Eq. (3.25) vanishes, as can be seen by integrating by parts.

Inserting now Eq. (3.23) for Λ_X into the stationarity equation (3.17) yields

$$0 = N - \mathbf{K}L - \mathbf{e^L}(N_X - \mathbf{I}_X \mathbf{K} L) = \left(\mathbf{I} - \mathbf{e^L} \mathbf{I}_X\right)(N - \mathbf{K}L). \tag{3.26}$$

Besides normalized solutions we are looking for, Eq. (3.26) also possesses solutions, possibly unnormalized, fulfilling $N = \mathbf{K}L$ for which Eq. (3.23) yields $\Lambda_X = 0$. This happens because we used Eq. (3.21) which is also valid for $\Lambda_X(x) = 0$. This is a solution for the Lagrange multiplier only for parameterizations of L where the normalization constraint is not active at a stationary point, where $\Lambda_X(x) = 0$ is possible for a normalized solution L^*. In that case normalization has to be checked for the stationary solution.

Defining

$$T_L = N - \mathbf{\Lambda_X} e^L, \tag{3.27}$$

the stationarity equation (3.17) acquires the form

$$\mathbf{K}L = T_L \tag{3.28}$$

which is in general a non-linear equation because T_L depends on L. Discretizing the function L, choosing a positive integer η and a positive-definite matrix $\mathbf{A}$, Eq. (3.28) can be solved by iteration

$$L^{(i+1)} = L^{(i)} + \eta \mathbf{A}^{-1} \left(T_L(L^{(i)}) - \mathbf{K}L^{(i)} \right), \tag{3.29}$$

starting from an initial guess $L^{(0)}$. Here the Lagrange multiplier Λ_X has to be adapted to fulfill condition (3.23) at the end of iteration. Iteration procedures will be discussed in more detail in Appendix C.

3.1.2 Normalization by parameterization: Error functional $E(g)$

Referring to the discussion in Section 2.1.6 we show that Eq. (3.26) can alternatively be obtained by ensuring normalization, instead of using Lagrange multipliers, explicitly by the parameterization

$$L(x,y) = g(x,y) - \ln \int dy'\, e^{g(x,y')}, \quad L = g - \ln Z_X, \tag{3.30}$$

and considering the functional

$$E(g) = -\langle N | g - \ln Z_X \rangle + \frac{1}{2} \langle g - \ln Z_X | \mathbf{K} | g - \ln Z_X \rangle . \tag{3.31}$$

The stationarity equation for $g(x,y)$ obtained by setting the functional derivative $\delta E(g)/\delta g$ to zero yields again Eq. (3.26). We check this, using

$$\frac{\delta \ln Z_X(x')}{\delta g(x,y)} = \delta(x - x') e^{L(x,y)}, \quad \frac{\delta \ln Z_X}{\delta g} = \mathbf{e^L I}_X = \left(\mathbf{I}_X \mathbf{e^L} \right)^T, \tag{3.32}$$

and

$$\mathbf{L}'(x,y;x',y') = \frac{\delta L(x',y')}{\delta g(x,y)} = \delta(x - x')\delta(y - y') - \delta(x - x') e^{L(x,y)}, \tag{3.33}$$

$$\mathbf{L}' = \mathbf{I} - \mathbf{e}^{\mathbf{L}}\mathbf{I}_X. \tag{3.34}$$

(Note that despite $\mathbf{I}_X = \mathbf{I}_X^T$ the expression $\mathbf{I}_X\mathbf{e}^{\mathbf{L}} \neq \mathbf{e}^{\mathbf{L}}\mathbf{I}_X = (\mathbf{I}_X\mathbf{e}^{\mathbf{L}})^T$ is not symmetric because $\mathbf{e}^{\mathbf{L}}$ depends on y and does not commute with the non-diagonal $\mathbf{I}_X$.) Hence we obtain as stationarity equation of functional $E(g)$, written in terms of $L(g)$, again Eq. (3.26)

$$0 = -\frac{\delta L}{\delta g}\frac{\delta E(g)}{\delta L} = \mathbf{L}'G_L = G_L - \mathbf{e}^{\mathbf{L}}\Lambda_X = \left(\mathbf{I} - \mathbf{e}^{\mathbf{L}}\mathbf{I}_X\right)(N - \mathbf{K}L), \tag{3.35}$$

where $G_L = N - \mathbf{K}L = -\delta E(g)/\delta L$ is the L–gradient of $-E(g)$. Working with g instead of L, Λ_X stands here directly for the L– and thus g–dependent expression (3.23) and is not interpreted as Lagrange multiplier function. (In the Lagrange multiplier approach used for L, the relation (3.23) is obtained by reference to the normalization condition.) Referring to the discussion following Eq. (3.26) we note that, when solving for g instead for L, unnormalized solutions fulfilling $N = \mathbf{K}L$ are not possible.

Thus we have seen how the implementation of normalization constraints by Lagrange multipliers and by explicit parameterization are equivalent. In contrast, changing integration variables from L to g in the full Bayesian predictive density results in the appearance of a Jacobian $\delta L/\delta g$ so that $p(L)dL = p(g)dg$. (To have an invertible Jacobian, for example, some value of g has to be fixed.) The presence of the Jacobian then modifies the MAP stationarity equations.

An alternative to $E(g)$ is to use a prior Gaussian in g instead in L, resulting in the functional

$$\tilde{E}(g) = -\langle N|\, g - \ln Z_X\,\rangle + \frac{1}{2}\langle g\,|\mathbf{K}|\, g\rangle, \tag{3.36}$$

where the normalization terms $\ln Z_X$, nonlocal in y, only appear in the likelihood term but, in contrast to $E(g)$, not in the prior term. The stationarity equation thus becomes

$$0 = N - \mathbf{K}g - \mathbf{e}^{\mathbf{L}}N_X = N - \mathbf{K}g - \mathbf{P}N_X. \tag{3.37}$$

In case $\ln Z_X$ is in the null space of $\mathbf{K}$, the Gaussian prior in L of functional $E(g)$ already corresponds to a Gaussian prior in g.

For density estimation problems with Gaussian specific priors, the solution g can be calculated in (X–)data space, i.e., in the space of functions defined on $Y \times x_D$ with $x_D = \{x_i | 1 \leq i \leq n\}$ representing the set of x–values of the training data. Let us denote by $\mathbf{I}_D$ the corresponding projector in data space $\mathbf{I}_D = \delta(y - y')\delta(x - x')\sum_{x_i=1}^{\tilde{n}}\delta(x - x_i)$, where $\tilde{n} \leq n$ denotes the number of different x_i–values. This is well defined for discrete variables where δ stands for the Kronecker-δ. For continuous variables, an integration of x_i over a finite interval I_i has to be included, $\delta(x - x_i) \to \int_{I_i} dx_i \delta(x - x_i)$, so that $\mathbf{I}_D^2$ becomes

well defined (when integrating over test functions) and equal to $\mathbf{I}_D$. Because of $\mathbf{I}_D N = N$, $\mathbf{I}_D \mathbf{P} N_X = \mathbf{P} N_X$, we obtain by projecting Eq. (3.37) on data space or its complement, respectively,

$$0 = N - \mathbf{K}_{DD}\, g_D - \mathbf{K}_{D\bar{D}}\, g_{\bar{D}} - \mathbf{P} N_X, \quad 0 = \mathbf{K}_{\bar{D}D}\, g_D + \mathbf{K}_{\bar{D}\bar{D}}\, g_{\bar{D}}, \tag{3.38}$$

where $\mathbf{I}_{\bar{D}} = \mathbf{I} - \mathbf{I}_D$, and $g_D = \mathbf{I}_D g$, $g_{\bar{D}} = \mathbf{I}_{\bar{D}} g$, $\mathbf{K}_{DD} = \mathbf{I}_D \mathbf{K} \mathbf{I}_D$, $\mathbf{K}_{D\bar{D}} = \mathbf{I}_D \mathbf{K} \mathbf{I}_{\bar{D}}$, $\mathbf{K}_{\bar{D}D} = \mathbf{I}_{\bar{D}} \mathbf{K} \mathbf{I}_D$, and $\mathbf{K}_{\bar{D}\bar{D}} = \mathbf{I}_{\bar{D}} \mathbf{K} \mathbf{I}_{\bar{D}}$. It follows that in data space (see, for example, [126], chapter III)

$$0 = N - \left(\mathbf{K}_{DD} - \mathbf{K}_{D\bar{D}} \mathbf{K}_{\bar{D}\bar{D}}^{-1} \mathbf{K}_{\bar{D}D}\right) g_D - \mathbf{P} N_X. \tag{3.39}$$

The same equation can be found by integrating a Gaussian specific prior factor $e^{-\langle g|\mathbf{K}|g\rangle/2}$ over the space of functions $g_{\bar{D}}$ defined on non-data points, leaving for data space functions g_D a factor proportional to

$$e^{-\frac{1}{2}\left\langle g_D \left| \mathbf{K}_{DD} - \mathbf{K}_{D\bar{D}} \mathbf{K}_{\bar{D}\bar{D}}^{-1} \mathbf{K}_{\bar{D}D} \right| g_D \right\rangle}. \tag{3.40}$$

Working in data space is especially useful in classification problems where y is discrete and only a summation but no integration over y is necessary. For Gaussian specific priors all calculations can then be done in a finite dimensional space [521].

3.1.3 The Hessians $\mathbf{H}_L$, $\mathbf{H}_g$

To design iteration procedures for solving stationarity equations it is useful to calculate the Hessians of the posterior or the energy functional. Besides their use in Newton-like iteration schemes (see Appendix C), the inverse Hessian calculated at the final solution also gives the error bars of the solution in a Gaussian approximation according to Eqs. (2.92, 2.93). Furthermore, knowledge of the Hessian is needed for calculating higher order contributions going beyond a saddle point approximation (MAP).

The Hessian $\mathbf{H}_L$ of $-E(L)$, Eq. (3.3), is given by

$$\mathbf{H}_L(L)(x,y;x',y') = \left.\frac{\delta^2(-E(L))}{\delta L(x,y)\delta L(x',y')}\right|_L. \tag{3.41}$$

For functional (3.3) and fixed Lagrange multiplier function Λ_X, we find the Hessian by taking the derivative of the gradient on the right of Eq. (3.17) with respect to L again. This gives

$$\mathbf{H}_L = -\mathbf{K} - \mathbf{\Lambda}_X \mathbf{e}^{\mathbf{L}} = -\mathbf{K} - \mathbf{\Lambda}_X \mathbf{P}. \tag{3.42}$$

The appearance of the diagonal matrix $\mathbf{\Lambda}_X \mathbf{e}^{\mathbf{L}} = \mathbf{e}^{\mathbf{L}} \mathbf{\Lambda}_X$ in Eq. (3.42) can result in a negative definite $\mathbf{H}$ even if $\mathbf{K}$ has zero modes, as in the case of a

differential operator $\mathbf{K}$ with periodic boundary conditions. Note that $\mathbf{\Lambda}_X \mathbf{e}^{\mathbf{L}}$ is diagonal and therefore symmetric, but not necessarily positive definite, because $\Lambda_X(x)$ can be negative for some x. Depending on the sign of $\Lambda_X(x)$, the equality constraint $Z_X(x) = 1$ for x can be replaced by the inequality constraint $Z_X(x) \leq 1$ or $Z_X(x) \geq 1$, respectively [40].

Including the L–dependence of Λ_X and with

$$\frac{\delta e^L}{\delta g} = \frac{\delta P}{\delta g} = \left(\mathbf{I} - \mathbf{e}^{\mathbf{L}} \mathbf{I}_X\right) \mathbf{e}^{\mathbf{L}} = \mathbf{e}^{\mathbf{L}} - \mathbf{e}^{\mathbf{L}} \mathbf{I}_X \mathbf{e}^{\mathbf{L}} = \mathbf{P} - \mathbf{P}\mathbf{I}_X\mathbf{P}, \tag{3.43}$$

we find, written in terms of L,

$$\mathbf{H}_g(L)(x,y;x',y') = \left.\frac{\delta^2(-E(g))}{\delta g(x,y)\delta g(x',y')}\right|_L$$

$$\begin{aligned}
= \; & -\mathbf{K}(x,y;x',y') - e^{L(x,y)} e^{L(x',y')} \int dy''dy''' \, \mathbf{K}(x,y'';x',y''') \\
& + e^{L(x',y')} \int dy'' \, \mathbf{K}(x,y;x',y'') + e^{L(x,y)} \int dy'' \, \mathbf{K}(x,y'';x',y') \\
& - \delta(x-x')\delta(y-y') e^{L(x,y)} \left(N_X(x) - \int dy'' \, (\mathbf{K}L)(x,y'') \right) \\
& + \delta(x-x') e^{L(x,y)} e^{L(x',y')} \left(N_X(x) - \int dy'' \, (\mathbf{K}L)(x,y'') \right),
\end{aligned} \tag{3.44}$$

The last term, diagonal in X, has dyadic structure in Y, and therefore for fixed x at most one non–zero eigenvalue. In matrix notation the Hessian reads

$$\begin{aligned}
\mathbf{H}_g &= -\left(\mathbf{I} - \mathbf{e}^{\mathbf{L}}\mathbf{I}_X\right) \mathbf{K} \left(\mathbf{I} - \mathbf{I}_X \mathbf{e}^{\mathbf{L}}\right) - \left(\mathbf{I} - \mathbf{e}^{\mathbf{L}}\mathbf{I}_X\right) \mathbf{\Lambda}_X \mathbf{e}^{\mathbf{L}} \\
&= -\left(\mathbf{I} - \mathbf{P}\mathbf{I}_X\right) \left[\mathbf{K}\left(\mathbf{I} - \mathbf{I}_X\mathbf{P}\right) + \mathbf{\Lambda}_X \mathbf{P}\right].
\end{aligned} \tag{3.45}$$

The expression is symmetric under $x \leftrightarrow x', y \leftrightarrow y'$, as required for a Hessian. This can be verified using the symmetry of $\mathbf{K} = \mathbf{K}^T$ and the fact that $\mathbf{\Lambda}_X$ and $\mathbf{I}_X$ commute. Because functional $E(g)$ is invariant under a shift transformation, $g(x,y) \to g(x,y) + c(x)$, the Hessian has a space of zero modes with the dimension of X. Indeed, any y–independent function (which can have finite L^1–norm only in finite Y–spaces) is a left eigenvector of $\left(\mathbf{I} - \mathbf{e}^{\mathbf{L}}\mathbf{I}_X\right)$ with eigenvalue zero. Thus, if necessary, the pseudo inverse has to be used instead of the inverse. Alternatively, additional constraints on g can be added which remove zero modes, as in boundary conditions.

3.2 Gaussian prior factor for likelihoods

3.2.1 Lagrange multipliers: Error functional $E(P)$

In the previous section we have formulated density estimation in terms of log–likelihoods as fields $\phi = L$, and accordingly prior factors have been cho-

sen Gaussian in terms of the log–likelihood. This is convenient because the positivity condition is then already implemented implicitly. The discussion of noise induced priors in 2.3.2, however, also provides arguments to choose prior factors which are Gaussian in terms of the likelihood $\phi(x,y) = P(x,y) = p(y|x,\phi)$. Hence we consider now a regularizing term which is quadratic in P instead of L and which corresponds to a Gaussian specific prior factor in P, so the posterior reads

$$p(\phi|f) \propto$$

$$e^{\sum_i \ln P(x_i,y_i) - \frac{1}{2}\int dx\,dy\,dx'\,dy'\,P(x,y)\mathbf{K}(x,y;x',y')P(x',y') + \int dx\,\Lambda_X(x)\left(1-\int dy\,P(x,y)\right)}. \tag{3.46}$$

Hence the relevant error functional becomes

$$E(P) = \beta E_{\text{comb}} = -\langle \ln P|\, N\rangle + \frac{1}{2}\langle P\,|\mathbf{K}|\,P\rangle + \langle P-\delta(y)|\,\Lambda_X\rangle\,. \tag{3.47}$$

Typical choices for $\mathbf{K}$ include differential operators preferring smooth solutions for P. The choice $\mathbf{K} = \lambda\mathbf{I}$, which corresponds to a regularization term

$$\frac{\lambda}{2}\langle P|\,P\rangle = \frac{\lambda}{2}||P||^2, \tag{3.48}$$

can be interpreted as a smoothness prior with respect to the distribution function of P (see Section 3.5).

In functional (3.47) we have implemented only the normalization constraint for the likelihood P by Lagrange multipliers, but not the non–negativity constraint. This is sufficient if $P(x,y) > 0$ for the stationary solution because then non–negativity holds for the likelihood also in some neighborhood and there are no components of the gradient pointing into regions with a negative likelihood. Hence in that case the non–negativity constraint is not active at the stationarity point. As probabilities or densities have to be positive at data points, smoothness constraints result typically in positive probabilities or densities everywhere unless set to zero explicitly by boundary conditions. If, however, the stationary solution possesses locations with $P(x,y) = 0$ at non–boundary points, then the component of the gradient pointing in the region with negative probabilities or densities has to be eliminated by introducing Lagrange parameters for each $P(x,y)$. This may happen if the regularizer rewards oscillatory behavior.

The stationarity equation for $E(P)$ of Eq. (3.47) is easily found

$$0 = \mathbf{P}^{-1}N - \mathbf{K}P - \Lambda_X, \tag{3.49}$$

with the diagonal matrix $\mathbf{P}$ defined in (3.8), or multiplied by $\mathbf{P}$

$$0 = N - \mathbf{P}\mathbf{K}P - \mathbf{P}\Lambda_X. \tag{3.50}$$

As likelihoods $P(x,y)$ can not be zero at observed data points (x_i,y_i), $\mathbf{P}^{-1}N$ is well defined.

Combining the normalization condition Eq. (3.21) for $\Lambda_X(x) \neq 0$ with Eq. (3.49) or (3.50) the Lagrange multiplier function Λ_X is found as

$$\Lambda_X = \mathbf{I}_X\,(N - \mathbf{P}\mathbf{K}P) = N_X - \mathbf{I}_X\mathbf{P}\mathbf{K}P, \tag{3.51}$$

where $\mathbf{I}_X\mathbf{P}\mathbf{K}P(x,y) = \int dy'dx''dy''\,P(x,y')\mathbf{K}(x,y';x'',y'')P(x'',y'')$. Eliminating Λ_X in Eq. (3.49) by using Eq. (3.51) gives finally

$$0 = (\mathbf{I} - \mathbf{I}_X\mathbf{P})(\mathbf{P}^{-1}N - \mathbf{K}P), \tag{3.52}$$

or for Eq. (3.50)

$$0 = (\mathbf{I} - \mathbf{P}\mathbf{I}_X)(N - \mathbf{P}\mathbf{K}P). \tag{3.53}$$

For similar reasons as discussed for Eq. (3.26), unnormalized solutions fulfilling $N = \mathbf{P}\mathbf{K}P$ are possible. Defining

$$T_P = \mathbf{P}^{-1}N - \Lambda_X = \mathbf{P}^{-1}N - N_X + \mathbf{I}_X\mathbf{P}\mathbf{K}P, \tag{3.54}$$

the stationarity equation can be written similar to Eq. (3.28) as

$$\mathbf{K}P = T_P, \tag{3.55}$$

with $T_P = T_P(P)$, and can be solved by iteration as in Eq. (3.29)

$$P^{(i+1)} = P^{(i)} + \eta\mathbf{A}^{-1}\left(T_P(P^{(i)}) - \mathbf{K}P^{(i)}\right), \tag{3.56}$$

starting from some initial guess $P^{(0)}$.

3.2.2 Normalization by parameterization: Error functional $E(z)$

Again, normalization can also be ensured by parameterization of P and solving for unnormalized probabilities or densities z, i.e.,

$$P(x,y) = \frac{z(x,y)}{\int dy\,z(x,y)}, \quad P = \frac{z}{Z_X}. \tag{3.57}$$

The corresponding error functional for z reads

$$E(z) = -\left\langle N \middle| \ln\frac{z}{Z_X} \right\rangle + \frac{1}{2}\left\langle \frac{z}{Z_X} \middle| \mathbf{K} \middle| \frac{z}{Z_X} \right\rangle. \tag{3.58}$$

We have

$$\frac{\delta z}{\delta z} = \mathbf{I}, \quad \frac{\delta Z_X}{\delta z} = \mathbf{I}_X, \quad \frac{\delta \ln z}{\delta z} = \mathbf{z}^{-1} = (\mathbf{Z}_X\mathbf{P})^{-1}, \quad \frac{\delta \ln Z_X}{\delta z} = \mathbf{Z}_X^{-1}\mathbf{I}_X, \tag{3.59}$$

with diagonal matrix $\mathbf{z}$ built analogous to $\mathbf{P}$ and $\mathbf{Z}_X$, and

$$\frac{\delta P}{\delta z} = \mathbf{P}' = \mathbf{Z}_X^{-1}\left(\mathbf{I} - \mathbf{I}_X\mathbf{P}\right), \quad \frac{\delta \ln P}{\delta z} = \mathbf{Z}_X^{-1}\left(\mathbf{P}^{-1} - \mathbf{I}_X\right), \tag{3.60}$$

$$\frac{\delta Z_X^{-1}}{\delta z} = -\mathbf{Z}_X^{-2}\mathbf{I}_X, \quad \frac{\delta P^{-1}}{\delta z} = -\mathbf{Z}_X^{-1}\left(\mathbf{I} - \mathbf{I}_X\mathbf{P}\right)\mathbf{P}^{-2}. \tag{3.61}$$

The diagonal matrices $\mathbf{Z}_X$, $\mathbf{P}$ commute, as well as $[\mathbf{Z}_X, \mathbf{I}_X] = 0$, but $[\mathbf{P}, \mathbf{I}_X] \neq 0$. Furthermore $(\mathbf{I} - \mathbf{P}\mathbf{I}_X) = (\mathbf{I} - \mathbf{I}_X\mathbf{P})^T$. Setting the gradient to zero we find

$$\begin{aligned} 0 = -\frac{\delta P}{\delta z}\frac{\delta E(z)}{\delta P} &= \mathbf{P}' G_P = \mathbf{Z}_X^{-1}\left(\mathbf{I} - \mathbf{I}_X\mathbf{P}\right)\left(\mathbf{P}^{-1}N - \mathbf{K}P\right) \\ &= \mathbf{Z}_X^{-1}\left(G_P - \Lambda_X\right) = \left(\mathbf{G}_P - \mathbf{\Lambda}_X\right)Z_X^{-1}, \end{aligned} \tag{3.62}$$

with P-gradient $G_P = \mathbf{P}^{-1}N - \mathbf{K}P = -\delta E(z)/\delta P$ of $-E(z)$ and $\mathbf{G}_P$ the corresponding diagonal matrix. Multiplied by $\mathbf{Z}_X$ Eq. (3.62) results in the stationarity equation (3.52), equivalent to (3.55).

3.2.3 The Hessians $\mathbf{H}_P$, $\mathbf{H}_z$

We now calculate the Hessian of the functional $-E(P)$. For fixed Lagrange multiplier function Λ_X one finds the Hessian by differentiating again the gradient of $-E(P)$ given on the right hand side of (3.49),

$$\mathbf{H}_P(P)(x,y;x'y') = -\mathbf{K}(x,y;x'y') - \delta(x-x')\delta(y-y')\sum_i \frac{\delta(x-x_i)\delta(y-y_i)}{P^2(x,y)}, \tag{3.63}$$

i.e.,

$$\mathbf{H}_P = -\mathbf{K} - \mathbf{P}^{-2}\mathbf{N}. \tag{3.64}$$

Here the diagonal matrix $\mathbf{P}^{-2}\mathbf{N}$ is non–zero only at data points.

Including now the dependence of Λ_X on P one obtains for the Hessian of $-E(z)$ in (3.58) by calculating the derivative of the gradient in (3.62)

$$\begin{aligned} \mathbf{H}_z(x,y;x',y') = -\frac{1}{Z_X(x)}\Big[&\mathbf{K}(x,y;x',y') \\ &- \int dy''\Big(P(x,y'')\mathbf{K}(x,y'';x',y') + \mathbf{K}(x,y;x',y'')P(x',y'')\Big) \\ &+ \int dy''dy'''P(x,y'')\mathbf{K}(x,y'';x',y''')P(x',y''') \\ &+\delta(x-x')\delta(y-y')\sum_i \frac{\delta(x-x_i)\delta(y-y_i)}{P^2(x,y)} - \delta(x-x')\sum_i \delta(x-x_i) \end{aligned}$$

$$-\delta(x-x')\int dx''dy''\Big(\mathbf{K}(x,y;x'',y'')P(x'',y'')+P(x'',y'')\mathbf{K}(x'',y'';x',y')\Big)$$

$$+2\,\delta(x-x')\int dy''dx'''dy'''P(x,y'')\mathbf{K}(x,y'';x''',y''')P(x''',y''')\Big]\frac{1}{Z_X(x')}, \tag{3.65}$$

i.e.,

$$\begin{aligned}\mathbf{H}_z &= \mathbf{Z}_X^{-1}\left(\mathbf{I}-\mathbf{I}_X\mathbf{P}\right)\left(-\mathbf{K}-\mathbf{P}^{-2}\mathbf{N}\right)\left(\mathbf{I}-\mathbf{P}\mathbf{I}_X\right)\mathbf{Z}_X^{-1}\\ &\quad -\mathbf{Z}_X^{-1}\left(\mathbf{I}_X\left(\mathbf{G}_P-\mathbf{\Lambda}_X\right)+\left(\mathbf{G}_P-\mathbf{\Lambda}_X\right)\mathbf{I}_X\right)\mathbf{Z}_X^{-1}, \qquad (3.66)\\ &= -\mathbf{Z}_X^{-1}\Big[\left(\mathbf{I}-\mathbf{I}_X\mathbf{P}\right)\mathbf{K}\left(\mathbf{I}-\mathbf{P}\mathbf{I}_X\right)+\mathbf{P}^{-2}\mathbf{N}\\ &\quad -\mathbf{I}_X\mathbf{P}^{-1}\mathbf{N}-\mathbf{N}\mathbf{P}^{-1}\mathbf{I}_X+\mathbf{I}_X\mathbf{N}\mathbf{I}_X\\ &\quad +\mathbf{I}_X\mathbf{G}_P+\mathbf{G}_P\mathbf{I}_X-2\,\mathbf{I}_X\mathbf{\Lambda}_X\Big]\mathbf{Z}_X^{-1}. \qquad (3.67)\end{aligned}$$

Here we used $[\mathbf{\Lambda}_X,\mathbf{I}_X]=0$. Like for (3.44), it follows from the normalization $\int dy\,P(x,y)=1$ that any y–independent function is right eigenvector of $(\mathbf{I}-\mathbf{I}_X\mathbf{P})$ with zero eigenvalue. With $\Lambda_X=\mathbf{I}_X\mathbf{P}G_P$ this factor or its transpose is also contained in the second line of Eq. (3.66), which means that $\mathbf{H}_z$ has a zero mode. Indeed, functional $E(z)$ is invariant under multiplication of z with a y–independent factor. The zero modes can be projected out or removed by including additional conditions, e.g., by fixing one value of the function $z(x,y)$ for every x.

3.3 Quadratic density estimation and empirical risk minimization

When interpreting an energy or error functional E probabilistically by assuming $-\beta E+c$ to be the logarithm of a posterior, the form of the training data term has to be $-\sum_i \ln P(x_i,y_i)$ like in Eqs. (3.3) or (3.47). Technically, however, it would be easier to replace such a data term by one which is quadratic in the function ϕ of interest.

Indeed, we have already mentioned in Section 2.2.5 that such functionals can be justified within the framework of empirical risk minimization. From that Frequentist point of view an error functional $E(P)$ is not derived from a log–posterior, but represents an empirical risk $\hat{r}(P,f)=\sum_i l(x_i,y_i,P)$, approximating an expected risk $r(P,f)$ for action $a=P$. This is possible under the assumption that training data are sampled according to the true $p(x,y|f)$. In that interpretation one is not restricted to a log–loss for training data but may as well choose for training data a quadratic loss like

$$E_D=\frac{1}{2}\left\langle\phi-\phi_{\rm emp}\,|\mathbf{K}_D|\,\phi-\phi_{\rm emp}\right\rangle, \tag{3.68}$$

with real–symmetric, positive (semi–)definite $\mathbf{K}_D$ and empirical reference ϕ_{emp}, consistent with the empirical density $P_{\text{emp}} = P(\phi_{\text{emp}})$ representing the training data. In particular, for $\phi = P$ we have

$$E_D = \langle P - P_{\text{emp}} \, |\mathbf{K}_D| \, P - P_{\text{emp}} \rangle \, /2. \tag{3.69}$$

When reconstructing (approximating) a joint probability density $p(x, y|\phi)$, the reference density P_{emp} is the joint empirical density

$$P_{\text{emp}}(x, y) = \frac{1}{n} \sum_i^n \delta(x - x_i)\delta(y - y_i) = \frac{N(x, y)}{n}, \tag{3.70}$$

with N of (3.4) as obtained from the training data. For reconstruction of a conditional probability density $p(y|x, \phi)$ the reference P_{emp} has to be chosen as conditional empirical density,

$$P_{\text{emp}}(x, y) = \frac{\sum_i \delta(x - x_i)\delta(y - y_i)}{\sum_i \delta(x - x_i)} = \frac{N(x, y)}{n_x}. \tag{3.71}$$

Eq. (3.71) is only valid if $n_x \neq 0$. For x–values not appearing in the training data the conditional empirical density (3.71) is not defined. (In contrast, the conditional empirical density for y–values which do not appear as training data is zero and thus well defined, if the corresponding x appears in the training data.) For x–variables with a large number of possible values, it can not be assumed that all x appear in the training data. For continuous x–variables this is even impossible. Hence, when approximating conditional densities P, either non–data x–values must be excluded from the integration in (3.68) by using an operator $\mathbf{K}_D$ which is restricted to x appearing as training data, or P_{emp} must be defined also for such non–data x–values. For existing $V_X = \mathbf{I}_X 1 = \int dy \, 1$, a possible extension $\tilde{P}_{\text{emp}}$ of P_{emp} is to assume a uniform density for non–data x–values, yielding

$$\tilde{P}_{\text{emp}}(x, y) = \begin{cases} \frac{\sum_i \delta(x - x_i)\delta(y - y_i)}{\sum_i \delta(x - x_i)} & \text{for all } x \text{ in the training data,} \\ \frac{1}{\int dy \, 1} & \text{for all } x \text{ not in the training data.} \end{cases} \tag{3.72}$$

This introduces a bias towards uniform densities, but has the advantage to give a conditional empirical density for all x and to fulfill the conditional normalization requirements.

Instead of a quadratic term in P, one can choose a quadratic term in another function ϕ, like a log–likelihood L. As every empirical log–likelihood is minus infinity at all non–data points $(x, y) \notin D$, one can instead work with a finite expression by choosing a small $\epsilon(y)$ and approximate P_{emp} by

$$P^{\epsilon}_{\text{emp}}(x, y) = \frac{\epsilon(y) + \sum_i \delta(x - x_i)\delta(y - y_i)}{\int dy \, \epsilon(y) + \sum_i \delta(x - x_i)}, \tag{3.73}$$

provided $\int dy\, \epsilon(y)$ exists. For $\epsilon(y) \neq 0$ also $P^{\epsilon}_{\text{emp}}(x,y) \neq 0$ for all x and $L^{\epsilon}_{\text{emp}} = \ln P^{\epsilon}_{\text{emp}} > -\infty$ exists.

Assuming a Gaussian prior factor, a quadratic data term in P results in an error functional

$$\tilde{E}(P) = \frac{1}{2}\langle P - P_{\text{emp}} |\mathbf{K}_D| P - P_{\text{emp}}\rangle + \frac{1}{2}\langle P |\mathbf{K}| P\rangle + \langle P|\Lambda_X\rangle, \qquad (3.74)$$

skipping the constant part of the Λ_X–terms. If required by $\mathbf{K}_D$, the empirical density P_{emp} in (3.74) can be replaced by a $\tilde{P}_{\text{emp}}$ defined for all x as in (3.72).

Positive (semi–)definite operators $\mathbf{K}_D$ have a square root and can be written in the form $\mathbf{R}^T\mathbf{R}$. One possibility, skipping for the sake of simplicity x in the following, is to choose as square root $\mathbf{R}$ the integration operator, $\mathbf{R} = \prod_k \mathbf{R}_k$ and $\mathbf{R}(y,y') = \theta(y-y')$. Thus $\mathbf{R}$ transforms the density function P into the distribution function ϕ. For a d_y–dimensional y–variable and a d_x–dimensional x–variable the integration operator $\mathbf{R} = \prod_{k=1}^{d_y} \mathbf{R}_k$ has matrix elements

$$\begin{aligned}\mathbf{R}(x,y;x',y') &= \prod_{k=1}^{d_y} \mathbf{R}_k(x,y;x',y') \\ &= \prod_{l=1}^{d_x} \delta(x_l - x'_l) \prod_{k=1}^{d_y} \theta(y_k - y'_k) \\ &= \delta(x-x')\theta(y-y') \qquad (3.75)\end{aligned}$$

and $\mathbf{R}_k$

$$\mathbf{R}_k(x,y;x',y') = \delta(x-x')\theta(y_k - y'_k) \prod_{l\neq k} \delta(y_l - y'_l). \qquad (3.76)$$

Hence the field $\phi = \mathbf{R}P$, where the product notation is understood as defined in (3.12), becomes

$$\phi(x,y) = \int_{-\infty}^{\infty} dx'\, dy'\, R(x,y,x',y')\, P(x',y') = \int_{-\infty}^{y} dy'\, P(x,y'). \qquad (3.77)$$

Analogously, the density P is recovered from the distribution function ϕ by differentiation. Thus for the inverse mapping we can write

$$P(\phi) = \prod_{k=1}^{d_y} \frac{\partial \phi}{\partial y_k} = \prod_{k=1}^{d_y} \nabla_{y_k}\phi = \prod_{k=1}^{d_y} \mathbf{R}_k^{-1}\phi. = \mathbf{R}^{-1}\phi, \qquad (3.78)$$

where $\mathbf{R}^{-1}$ is the differentiation operator $\prod_k \mathbf{R}_k^{-1} = \prod_k \nabla_{y_k}$ (with appropriate boundary conditions). Hence $\left(\mathbf{R}^T\right)^{-1}\mathbf{R}^{-1} = -\prod_k \Delta_k$ is the product of one–dimensional Laplacians $\Delta_k = \partial^2/\partial y_k^2$.

As already mentioned the normalization and non–negativity conditions for P appear for ϕ as boundary and monotonicity conditions. The boundary conditions can be implemented explictly and the monotonicity condition by Lagrange multipliers. Similarly to the positivity condition for P, the Lagrange multiplier can be skipped in practice if the ϕ which minimizes an error functional is everywhere non–constant because then the gradient at the stationary point has no components pointing into the forbidden area of nonmonotone distribution functions. (Even if the Lagrange multiplier can be skipped in such cases, the monotony condition has to be fulfilled, which thus must be enforced during iteration and checked for the final solution ϕ.) Thus skipping the Λ_X–term and adding to (3.74), for example, a regularizing term $\frac{\lambda}{2}\langle P|\,P\rangle$ as in (3.48) gives

$$\tilde{E}(P) = \frac{1}{2}\left\langle P - P_{\text{emp}}\left|\mathbf{R}^T\mathbf{R}\right|P - P_{\text{emp}}\right\rangle + \frac{\lambda}{2}\langle P|\,P\rangle \tag{3.79}$$

$$= \frac{1}{2}\left(\langle\phi - \phi_{\text{emp}}|\,\phi - \phi_{\text{emp}}\rangle - \lambda\left\langle\phi\left|\prod_k \Delta_k\right|\phi\right\rangle\right) \tag{3.80}$$

$$= \frac{1}{2m^2}\left\langle\phi\left|-\prod_k \Delta_k + m^2\mathbf{I}\right|\phi\right\rangle - \langle\phi|\,\phi_{\text{emp}}\rangle + \frac{1}{2}\langle\phi_{\text{emp}}|\,\phi_{\text{emp}}\rangle\,, \tag{3.81}$$

with $m^2 = \lambda^{-1}$. Here the empirical distribution function $\phi_{\text{emp}} = \mathbf{R}P_{\text{emp}}$ is given by $\phi_{\text{emp}}(y) = \frac{1}{n}\sum_i \theta(y - y_i)$ or, including the x–variable, $\phi_{\text{emp}}(x, y) = \frac{1}{N_X(x)}\sum_{x' \in x_D} \delta(x - x')\theta(y - y_i)$ for $N_X(x) \neq 0$, with $N_X(x)$ as defined in Eq. (3.24). This can be extended to a linear $\tilde{\phi} = \mathbf{R}\tilde{P}_{\text{emp}}$ for $N_X(x) = 0$. The stationarity condition yields[1]

$$\phi = m^2\left(-\prod_k \Delta_k + m^2\mathbf{I}\right)^{-1}\phi_{\text{emp}}. \tag{3.82}$$

Kernel methods of density estimation, like the use of Parzen windows, can be founded on such quadratic regularization functionals [492]. Indeed, a one–dimensional equation of the form (3.82) is equivalent to the use of Parzen's kernel in density estimation [361,382].

3.4 Numerical case study: Density estimation with Gaussian specific priors

In this section we present some basic numerical examples of nonparametric density estimation comparing the two choices for the field $\phi = L$ and $\phi =$

[1]For $d_y = 1$ (or $\phi = \prod_k \phi_k$) the operator in (3.82) becomes $\left(-\Delta + m^2\mathbf{I}\right)^{-1}$, which has the structure of a free massive propagator for a scalar field with mass m^2 and is calculated in Appendix C.2.3.

P. To this end we consider a problem with one–dimensional X–space, one–dimensional Y–space, and a smoothness prior with inverse covariance

$$\mathbf{K} = \lambda_x \left(\mathbf{K}_X \otimes 1_Y\right) + \lambda_y \left(1_X \otimes \mathbf{K}_Y\right), \tag{3.83}$$

where

$$\mathbf{K}_X = \lambda_0 \mathbf{I}_X - \lambda_2 \Delta_x + \lambda_4 \Delta_x^2 - \lambda_6 \Delta_x^3 \tag{3.84}$$

$$\mathbf{K}_Y = \lambda_0 \mathbf{I}_Y - \lambda_2 \Delta_y + \lambda_4 \Delta_y^2 - \lambda_6 \Delta_y^3, \tag{3.85}$$

and Laplacian

$$\Delta_x(x,x') = \delta''(x-x') = \delta(x-x')\frac{\partial^2}{\partial x^2}, \tag{3.86}$$

analogous for Δ_y. The derivative of the δ–function is defined by formal partial integration

$$\int_{-\infty}^{\infty} dy'\, f(y')\delta'(y-y') = f(y')\delta(y'-y)|_{-\infty}^{\infty} - f'(y). \tag{3.87}$$

For the special case $\lambda_2 \neq 0$, $\lambda_0 = \lambda_4 = \lambda_6 = 0$, this corresponds to the two Laplacian prior factors Δ_x for x and Δ_y for y. (Notice that even for $\lambda_x = \lambda_y$ the λ_4– and λ_6–terms do not include all terms of an iterated 2–dimensional Laplacian, like $\Delta^2 = (\Delta_x + \Delta_y)^2$ or Δ^4, as the mixed derivatives $\Delta_x \Delta_y$ are missing.)

We will now study nonparametric density estimation with Gaussian prior factors with respect to L as well as with respect to P. The error or energy functional $E(L)$ for a Gaussian prior factor in L is given by Eq. (3.2), the corresponding iteration procedure (3.29) being

$$L^{(i+1)} = L^{(i)} + \eta \mathbf{A}^{-1}\left(N - \mathbf{K}L^{(i)} - \mathbf{e}^{\mathbf{L}^{(i)}}\left[N_X - \mathbf{I}_X \mathbf{K} L^{(i)}\right]\right). \tag{3.88}$$

Written explicitly for the special case $\lambda_2 = 1$, $\lambda_0 = \lambda_4 = \lambda_6 = 0$, Eq. (3.88) reads,

$$\begin{aligned} L^{(i+1)}(x,y) &= L^{(i)}(x,y) + \eta \sum_j \mathbf{A}^{-1}(x,y;x_j,y_j) \\ &+\eta \int dx'dy' \mathbf{A}^{-1}(x,y;x',y') \left[\frac{\partial^2}{\partial(x')^2} L^{(i)}(x',y') + \frac{\partial^2}{\partial(y')^2} L^{(i)}(x',y')\right. \\ &-\left(\sum_j \delta(x'-x_j) + \int dy'' \frac{\partial^2}{\partial(x')^2} L^{(i)}(x',y'')\right. \\ &\left.\left. + \int dy'' \frac{\partial^2}{\partial(y'')^2} L^{(i)}(x',y'')\right) e^{L^{(i)}(x',y')}\right]. \end{aligned} \tag{3.89}$$

Here

$$\int_{y_A}^{y_B} dy'' \frac{\partial^2}{\partial (y'')^2} L^{(i)}(x', y'') = \frac{\partial}{\partial (y'')} L^{(i)}(x', y'') \Big|_{y_A}^{y_B}$$

vanishes on an interval $[y_A y_B]$, if the first derivative $\frac{\partial}{\partial y} L^{(i)}(x, y)$ vanishes at the boundary or is periodic.

For the error functional $E(P)$ of Eq. (3.47) the iteration procedure (3.56),

$$P^{(i+1)} = P^{(i)} + \eta \mathbf{A}^{-1} \left[(\mathbf{P}^{(i)})^{-1} N - N_X + \mathbf{I}_X \mathbf{P}^{(i)} \mathbf{K} P^{(i)} - \mathbf{K} P^{(i)} \right], \quad (3.90)$$

becomes for $\lambda_2 = 1$, $\lambda_0 = \lambda_4 = \lambda_6 = 0$

$$P^{(i+1)}(x, y) = P^{(i)}(x, y) + \eta \sum_j \frac{\mathbf{A}^{-1}(x, y; x_j, y_j)}{P^{(i)}(x_j, y_j)}$$

$$+\eta \int dx' dy' \mathbf{A}^{-1}(x, y; x', y') \left[\frac{\partial^2}{\partial (x')^2} P^{(i)}(x', y') + \frac{\partial^2}{\partial (y')^2} P^{(i)}(x', y') \right.$$

$$- \left(\sum_j \delta(x' - x_j) + \int dy'' P^{(i)}(x', y'') \frac{\partial^2 P^{(i)}(x', y'')}{\partial (x')^2} \right.$$

$$\left. \left. + \int dy'' P^{(i)}(x', y'') \frac{\partial^2 P^{(i)}(x', y'')}{\partial (y'')^2} \right) \right]. \quad (3.91)$$

Here

$$\int_{y_A}^{y_B} dy'' P^{(i)}(x', y'') \frac{\partial^2 P^{(i)}(x', y'')}{\partial (y'')^2} =$$

$$P^{(i)}(x', y'') \frac{\partial P^{(i)}(x', y'')}{\partial (y'')} \Big|_{y_A}^{y_B} - \int_{y_A}^{y_B} dy'' \left(\frac{\partial P^{(i)}(x', y'')}{\partial y''} \right)^2, \quad (3.92)$$

where the first term vanishes for $P^{(i)}$ periodic or vanishing at the boundaries.

The following numerical calculations have been performed on a mesh of dimension 10×15, $x \in [1, 10]$ and $y \in [1, 15]$, with periodic boundary conditions on y and sometimes also in x. Several different iteration and initialization methods have been used. Figs. 3.1 – 3.4 summarize results for density estimation problems with only two data points, where differences in the effects of varying smoothness priors are particularly easy to see. A density estimation with more data points is shown in Fig. 3.5.

For Fig. 3.1, a Laplacian smoothness prior on L has been implemented. The solution has been obtained by iterating with the negative Hessian, as long as it is positive definite (see Appendix C). Otherwise the gradient algorithm has been used. One iteration step means one iteration according to Eq. (3.29)

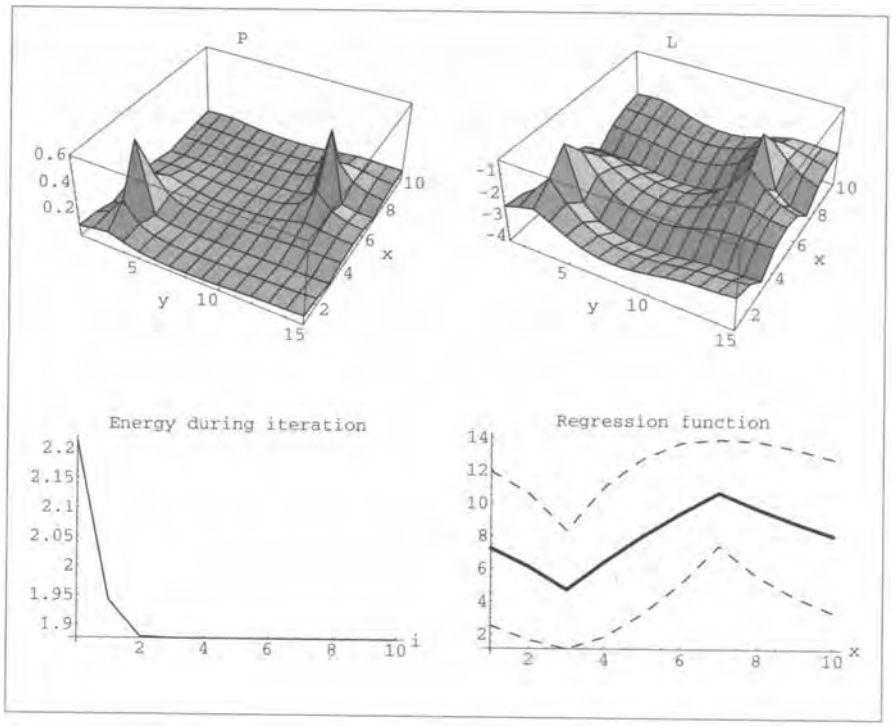


Figure 3.1: Density estimation 1. Density estimation with 2 data points and a Gaussian prior factor for the log–probability L. First row: Final P and L. Second row: The left hand side shows the energy $E(L)$ as defined in (3.3) during iteration, the right hand side shows the regression function $\phi(x) = \int dy\, y p(y|x,\phi) = \int dy\, y P(x,y)$. The dotted lines indicate the range of one standard deviation above and below the regression function (ignoring periodicity in x). The fast convergence to the minimal energy shows that the problem is nearly linear. The asymmetry of the solution between the x– and y–direction is due to the normalization constraint, only required for y. (Laplacian smoothness prior $\mathbf{K}$ as given in Eq. (3.83) with $\lambda_x = \lambda_y = 1$, $\lambda_0 = 0$, $\lambda_2 = 0.025$, $\lambda_4 = \lambda_6 = 0$. Iteration with negative Hessian $\mathbf{A} = -\mathbf{H}$ if positive definite, otherwise with the gradient algorithm, $\mathbf{A} = \mathbf{I}$. Initialization with $L^{(0)} = \ln(\tilde{\mathbf{C}}\tilde{P}_{\text{emp}})$, i.e., $L^{(0)}$ normalized to $\int dy\, e^L = 1$, with $\tilde{\mathbf{C}}$ of Eq. (C.88) and $\mathbf{C} = (\mathbf{K} + m_C^2\mathbf{I})$, $m_C^2 = 0.1$. Within each iteration step the optimal step width η has been found by a line search. Mesh with 10 points in x-direction and 15 points in y–direction, periodic boundary conditions in x and y. The 2 data points are $(3,3)$ and $(7,12)$.)

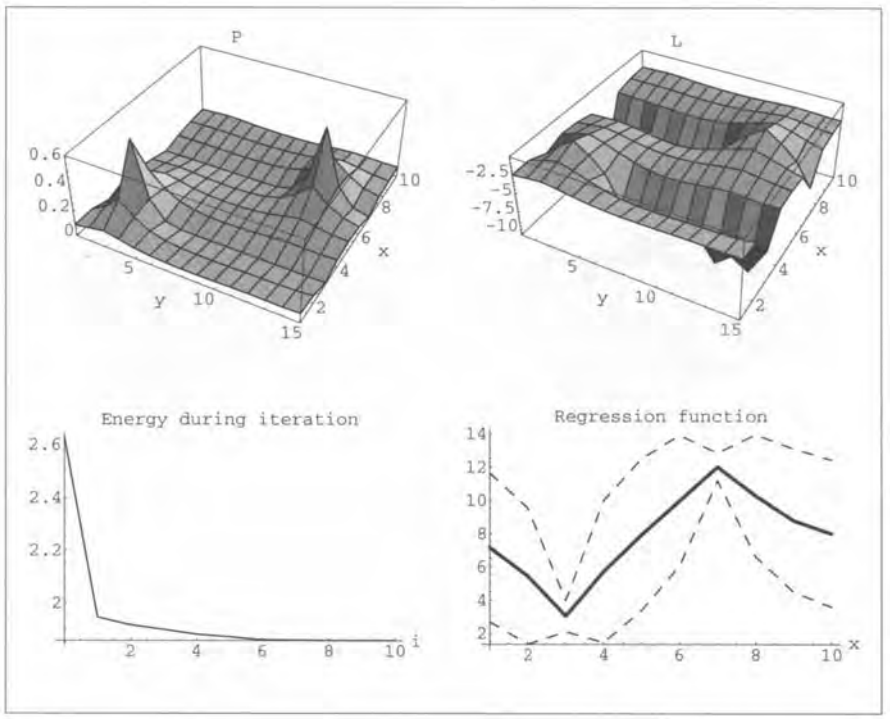


Figure 3.2: Density estimation 2. Density estimation with 2 data points, this time with a Gaussian prior factor for the probability P, minimizing the energy functional $E(P)$ as defined in (3.47). To make the figure comparable with Fig. 3.1, the parameters have been chosen so that the maximum of the solution P is the same in both figures ($\max P = 0.6$). Notice that, compared to Fig. 3.1, the smoothness prior is less effective for small probabilities. (Same data, mesh, and periodic boundary conditions as for Fig. 3.1. Laplacian smoothness prior $\mathbf{K}$ as in Eq. (3.83) with $\lambda_x = \lambda_y = 1$, $\lambda_0 = 0$, $\lambda_2 = 1$, $\lambda_4 = \lambda_6 = 0$. Iterated using massive prior relaxation, $\mathbf{A} = \mathbf{K} + m^2\mathbf{I}$ with $m^2 = 1.0$ (see Appendix C.1). Initialization with $P^{(0)} = \tilde{\mathbf{C}}\tilde{P}_{\text{emp}}$, with $\tilde{\mathbf{C}}$ of Eq. (C.88) so that $P^{(0)}$ is correctly normalized, and $\mathbf{C} = (\mathbf{K} + m_C^2\mathbf{I})$, $m_C^2 = 1.0$. Within each iteration step the optimal factor η has been found by a line search algorithm.)

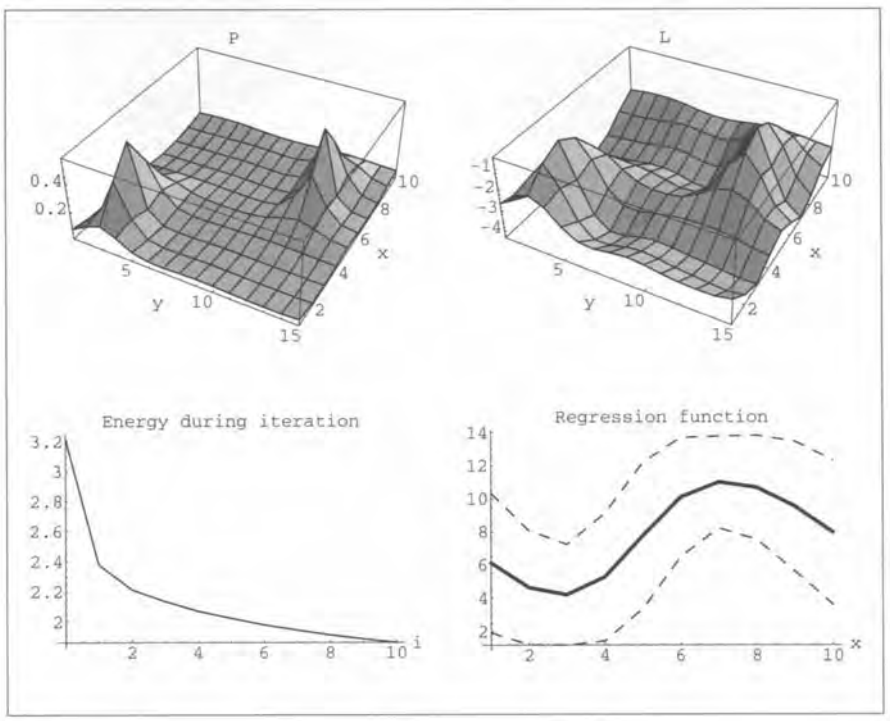


Figure 3.3: Density estimation 3. Density estimation with 2 data points with a $(-\Delta^3)$ Gaussian prior factor for the log–probability L. Such a prior favors probabilities of Gaussian shape. (Smoothness prior $\mathbf{K}$ of the form (3.83) with $\lambda_x = \lambda_y = 1$, $\lambda_0 = \lambda_2 = \lambda_4 = 0$, $\lambda_6 = 0.01$. Same iteration procedure, initialization, data, mesh, and periodic boundary conditions as for Fig. 3.1.)

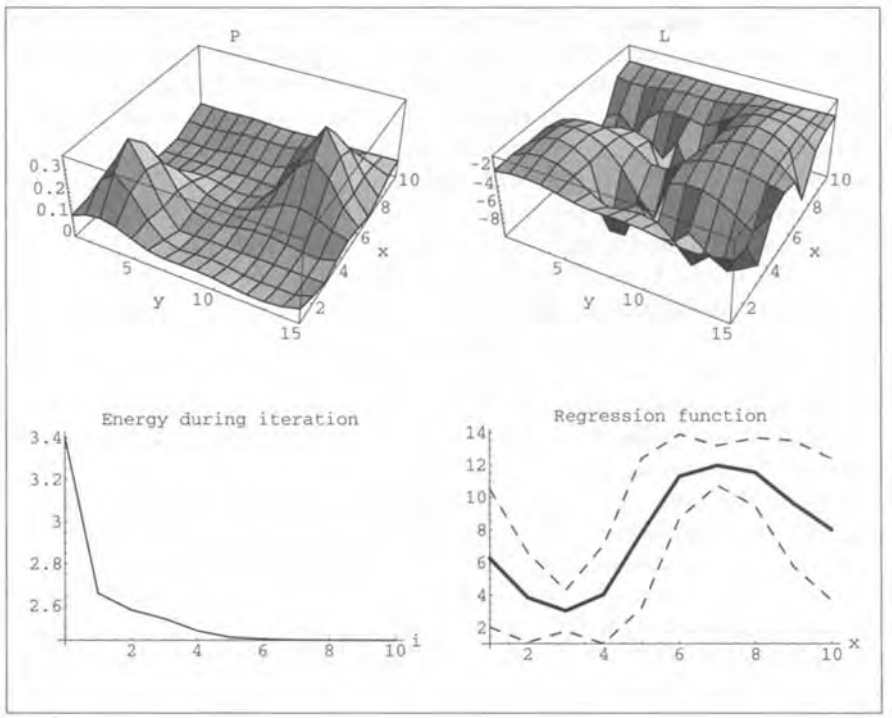


Figure 3.4: Density estimation 4. Density estimation with 2 data points with a $(-\Delta^3)$ Gaussian prior factor for the probability P. As the variation of P is smaller than that of L, a larger λ_6 has been chosen than in Fig. 3.3. The Gaussian prior in P is less effective for small probabilities than a comparable Gaussian prior in L. (Smoothness prior $\mathbf{K}$ of the form (3.83) with $\lambda_x = \lambda_y = 1$, $\lambda_0 = \lambda_2 = \lambda_4 = 0$, $\lambda_6 = 0.1$. Same iteration procedure, initialization, data, mesh, and periodic boundary conditions as for Fig. 3.2.)

with the optimal η. Thus each iteration step includes the optimization of η by a line search algorithm.

As initial guess in Fig. 3.1 the kernel estimate $L^{(0)} = \ln(\tilde{\mathbf{C}}\tilde{P}_{\text{emp}})$ has been employed. (See Eq. (C.85) of Appendix C. The kernel $\tilde{\mathbf{C}}$ is defined in Eq. (C.88) of Appendix C for $\mathbf{C} = (\mathbf{K} + m_C^2\mathbf{I})$ with squared mass $m_C^2 = 0.1$ as in Eq. (C.72).) The fast drop–off of the energy $E(L)$ within the first two iterations shows the quality of this initial guess. Indeed, this fast convergence seems to indicate that the problem is nearly linear, meaning that the influence of the only nonlinear term in the stationarity equation, the normalization constraint, is not too strong. Notice also that the reconstructed regression shows the typical piecewise linear approximations well known from one–dimensional (normalization constraint free) regression problems with Laplacian prior.

Fig. 3.2 shows a density estimation similar to Fig. 3.1, but for a Gaussian prior factor in P and thus with different λ_2, different initialization, and slightly different iteration procedure. For Fig. 3.2 a kernel estimate $P^{(0)} = (\tilde{\mathbf{C}}\tilde{P}_{\text{emp}})$ has been used as initial guess, again with $\tilde{\mathbf{C}}$ as defined in Eq. (C.88) of Appendix C and $\mathbf{C} = (\mathbf{K} + m_C^2\mathbf{I})$ but with squared mass $m_C^2 = 1.0$. The solution has been obtained by prior relaxation $\mathbf{A} = \mathbf{K} + m^2\mathbf{I}$ including a mass term with $m^2 = 1.0$ to get an invertible $\mathbf{A}$ for a Laplacian $\mathbf{K} = -\Delta$ and periodic boundary conditions, as discussed in Appendix C. This iteration scheme does not require to calculate the Hessian $\mathbf{H}_P$ at each iteration step. Again the quality of the initial guess (and the iteration scheme) is indicated by the relatively fast drop–off of the energy $E(P)$ during the first iteration.

Because the range of P–values is smaller than that of L–values, a larger Laplacian smoothness factor λ_2 is needed for Fig. 3.2 to get similar results as for Fig. 3.1. In particular, λ_2–values have been chosen for the two figures such that the maximal values of the two reconstructed probability densities P turn out to be nearly equal. (For practical problems, factors like the λ_i can, for example, be determined by cross–validation. For details see Chapter 5, in particular Section 5.4.5.)

Furthermore, because the logarithm expands the distances between small probabilities one expects a Gaussian prior for L to be especially effective for small probabilities. Comparing Fig. 3.1 and Fig. 3.2, this effect can indeed be seen. The deep valleys appearing in the L–landscape of Fig. 3.2 show that small values of L are not smoothed out as effectively as in Fig. 3.1. The variance of the solution $p(y|x, \phi)$ is much smaller for a Gaussian prior in P at those x which are in the training set.

Fig. 3.3 resumes the results for a model similar to that presented in Fig. 3.1, but with a Gaussian prior on L with inverse covariance $-\Delta^3$ replacing the Gaussian prior with Laplacian inverse covariance $-\Delta$. As all quadratic functions have zero third derivative, such a prior on L favors log–likelihoods which are quadratic in x, y, corresponding to likelihoods P, which are Gaussian in x, y. Indeed, this is indicated by the striking difference between the regression functions in Fig. 3.3 and in Fig. 3.1: The $-\Delta^3$–prior produces a much rounder

regression function, especially at the x–values which appear in the data. Note however, that in contrast to a pure Gaussian regression problem, in density estimation an additional non–quadratic normalization constraint is present.

In Fig. 3.4 a similar prior with inverse covariance $-\Delta^3$ has been applied, but this time Gaussian in P instead of L. In contrast to a $-\Delta^3$–prior for L, a $-\Delta^3$–prior for P implements a tendency to likelihoods (instead of log–likelihoods) which are quadratic in x, y. Similar to the difference between Fig. 3.1 and Fig. 3.3, the regression function in Fig. 3.4 is rounder than that in Fig. 3.2. Furthermore, smoothing in Fig. 3.4 is less effective for smaller probabilities than it is in Fig. 3.3. This is the same result we have found comparing the two priors for L shown in Fig. 3.2 and Fig. 3.1. Using the same λ_6 would lead, for a Gaussian prior on P, to deeper valleys in the L–landscape and to a smaller variance especially at locations x which appear in the training data.

Fig. 3.5 depicts the results of a density estimation based on more than two data points. Fifty training data have been obtained by sampling with uniform $p(x)$ from the 'true' mixture density, shown in the top row of Fig. 3.6,

$$P_{\text{true}}(x,y) = p(y|x,\phi_{\text{true}}) = \frac{1}{2\sqrt{2\pi}\sigma_0}\left(e^{-\frac{(y-\phi_a(x))^2}{2\sigma_0^2}} + e^{-\frac{(y-\phi_b(x))^2}{2\sigma_0^2}}\right), \quad (3.93)$$

with $\sigma_0 = 1.5$ and 'true' regression functions $\phi_a(x) = 125/18 + (5/9)x$ and $\phi_b(x) = 145/18 - (5/9)x$ for the two mixture components. The sampling process has been implemented using the transformation method (see, for example, [410]). The corresponding empirical density N/n of Eq. (3.70) and conditional empirical density P_{emp} of Eq. (3.71), in this case equal to the extended $\tilde{P}_{\text{emp}}$ defined in Eq. (3.72), can be found in Fig. 3.7.

Fig. 3.5 shows the maximum posterior solution $p(y|x,\phi^*)$ and its logarithm, the energy $E(L)$ during iteration, the *regression function*

$$\phi(x) = \int dy\, y\, p(y|x,\phi) = \int dy\, y\, P(x,y), \quad (3.94)$$

(as reference, the regression function for the true likelihood $p(y|x,\phi_{\text{true}})$ is given in Fig. 3.8), the *average training error* (or *empirical (conditional) log–loss*)

$$< -\ln p(y|x,\phi) >_D = -\frac{1}{n}\sum_{i=1}^{n} \ln p(y_i|x_i,\phi), \quad (3.95)$$

and the *average test error* (or *true expectation of (conditional) log–loss*) for uniform $p(x)$

$$< -\ln p(y|x,\phi) >_{P_{\text{true}}} = -\int dy\, dx\, p(x) p(y|x,\phi_{\text{true}}) \ln p(y|x,\phi), \quad (3.96)$$

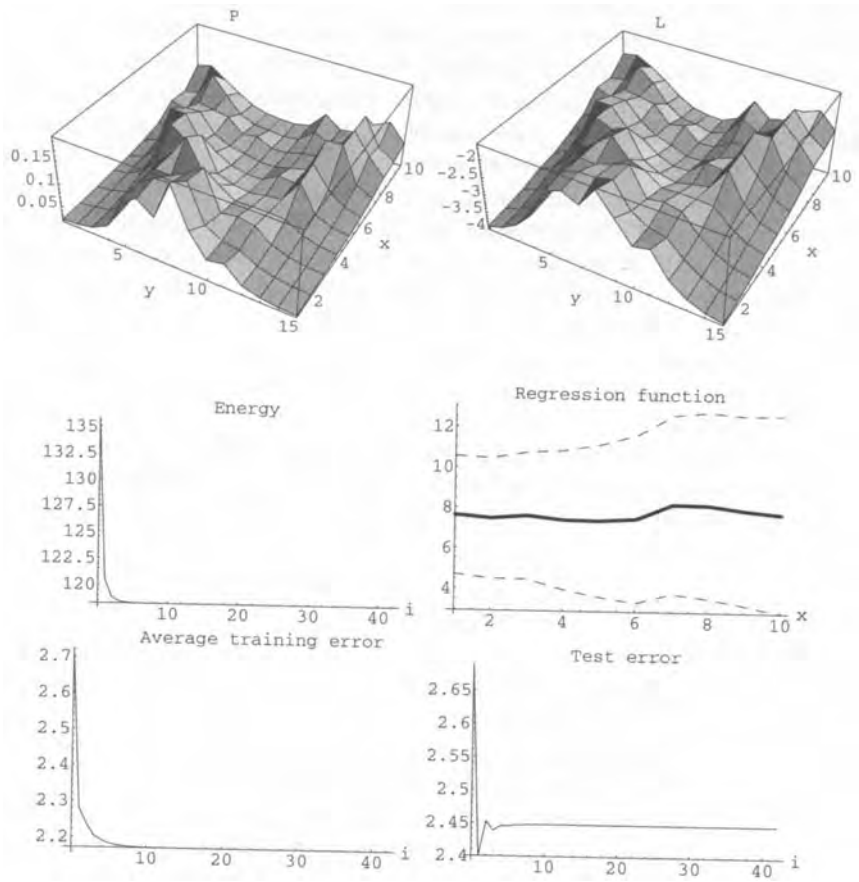


Figure 3.5: Density estimation 5. Density estimation with a Gaussian prior factor for the log–probability L and the 50 data points shown in Fig. 3.7. Top row: Final solution $P(x,y) = p(y|x,\phi)$ and $L = \ln P$. Second row: Energy $E(L)$ during iteration and final regression function. Bottom row: Average training error $-(1/n)\sum_{i=1}^{n} \ln p(y_i|x_i,\phi)$ during iteration and average test error $-\int dy\, dx\, p(x)p(y|x,\phi_{\text{true}}) \ln p(y|x,\phi)$ for uniform $p(x)$. Because the solution has been obtained from the training data, the test error is larger then the training error. (Parameters: Zero mean Gaussian smoothness prior with inverse covariance $\lambda\mathbf{K}$, $\lambda = 0.5$ and $\mathbf{K}$ of the form (3.83) with $\lambda_x = 2$, $\lambda_y = 1$, $\lambda_0 = 0$, $\lambda_2 = 1$, $\lambda_4 = \lambda_6 = 0$, massive prior iteration with $\mathbf{A} = \mathbf{K} + m^2\mathbf{I}$ and squared mass $m^2 = 0.01$. Initialized with normalized constant L. At each iteration step the factor η has been adapted by a line search algorithm. Mesh with 10 points in x-direction and 15 points in y–direction, periodic boundary conditions in y.)

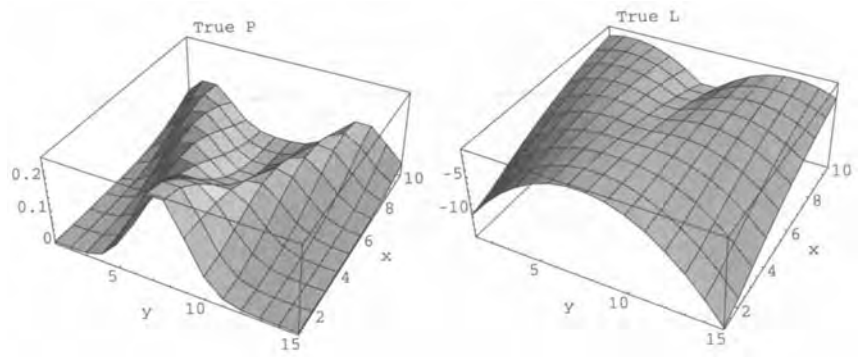


Figure 3.6: Density estimation 6. True density P_{true} (left hand side) and true log-density $L_{\text{true}} = \ln P_{\text{true}}$ (right hand side) used for Fig. 3.5 (and for Figs. 6.2 – 6.5 in Section 6.3). (As reference we give two values of the expected test error $\int dy\, dx\, p(x)p(y|x,\phi_{\text{true}}) \ln p(y|x,\phi)$ under the true $p(y|x,\phi_{\text{true}})$ for uniform $p(x)$: its minimal value is 2.23 for ϕ_{true}, for a uniform P it is 2.68.)

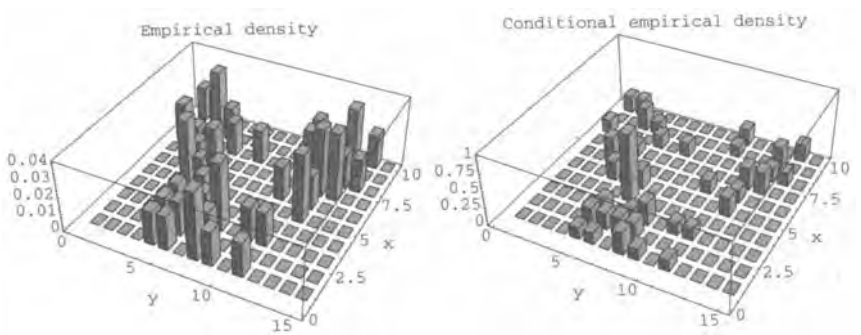


Figure 3.7: Empirical and conditional empirical density. Left hand side: Empirical density $N(x,y)/n = \sum_i \delta(x - x_i)\delta(y - y_i)/\sum_i 1$, sampled from $p(x,y|\phi_{\text{true}}) = p(y|x,\phi_{\text{true}})p(x)$ with uniform $p(x)$. Right hand side: Corresponding conditional empirical density $P_{\text{emp}}(x_i,y) = (\mathbf{N}_X^{-1}N)(x_i,y) = \sum_i \delta(x - x_i) \sum_i \delta(y - y_i)/p_{\text{emp}}(x_i)$. Both densities are obtained from the 50 data points used for Fig. 3.5 and for Figs. 6.2 – 6.5 in Section 6.3.

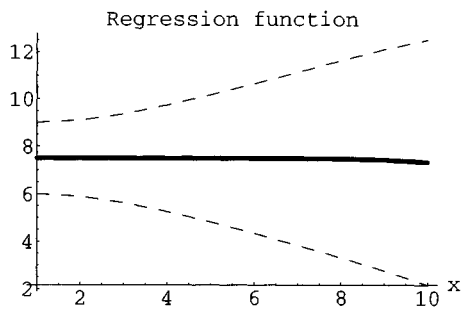


Figure 3.8: Regression function $\phi_{\text{true}}(x)$ for the true density P_{true} of Fig. 3.6, defined as $\phi(x) = \int dy\, y p(y|x, \phi_{\text{true}}) = \int dy\, y P_{\text{true}}(x, y)$. The dashed lines indicate the range of one standard deviation above and below the regression function.

which is, up to a constant, equal to the expected Kullback–Leibler distance between the actual solution and the true likelihood,

$$\text{KL}\Big(p(x, y|\phi_{\text{true}}), p(y|x, \phi)\Big) = -\int dy\, dx\, p(x, y|\phi_{\text{true}}) \ln \frac{p(y|x, \phi)}{p(y|x, \phi_{\text{true}})}. \quad (3.97)$$

The test error measures the quality of the achieved solution. It has, of course, in contrast to the energy and training error, not been available to the learning algorithm.

The maximum posterior solution of Fig. 3.5 has been calculated by minimizing $E(L)$ using massive prior iteration with $\mathbf{A} = \mathbf{K} + m^2\mathbf{I}$, a squared mass $m^2 = 0.01$, and as initial guess a constant $L^{(0)}$, normalized for each x. Convergence has been fast, the regression function is similar to the true one (see Fig. 3.8).

Fig. 3.9 compares different iteration procedures and initialization methods. Clearly, all methods do what they should do: they decrease the energy functional. Iterating with the negative Hessian yields the fastest convergence. Massive prior iteration is nearly as fast, even for uniform initialization, and does not require calculation of the Hessian at each iteration. The slowest iteration method, but easiest to implement, is the gradient algorithm.

Looking at Fig. 3.9 one can distinguish data–oriented from prior–oriented initializations. We understand data–oriented initial guesses to be those for which the training error is smaller at the beginning of the iteration than for the final solution and prior–oriented initial guesses to be those for which the opposite is true. Clearly, the uniform initialization is prior–oriented, while an empirical log–density $\ln(N/n + \epsilon)$ and the kernel initializations are data–oriented.

The case where the test error grows while the energy is decreasing indicates

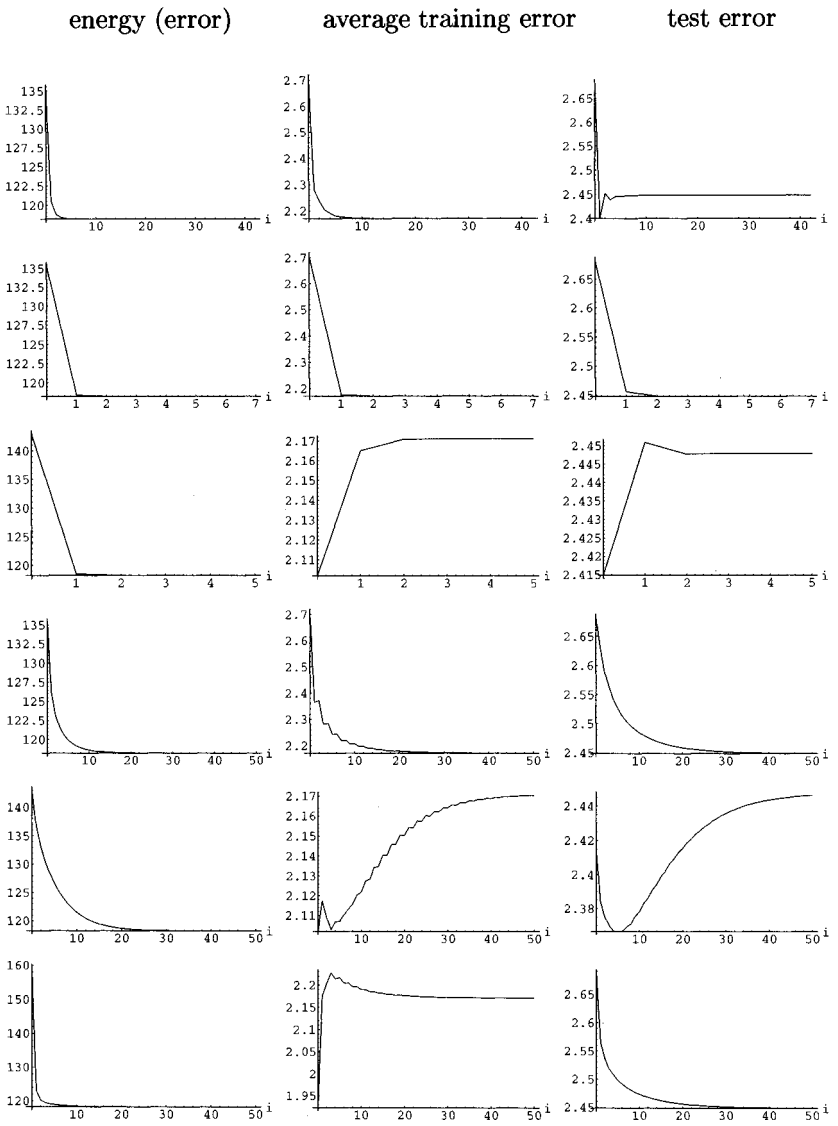


Figure 3.9: Comparison of iteration schemes and different initializations. Shown are the complete energy (error) $E(L)$, the corresponding average training error (as defined for Fig. 3.5), and the average test error with respect to the true solution (unknown to the algorithm, see Fig. 3.5) during iteration. First row: Massive prior iteration (with $\mathbf{A} = \mathbf{K} + m^2\mathbf{I}$, $m^2 = 0.01$) and uniform initialization. Second row: Hessian iteration ($\mathbf{A} = -\mathbf{H}$) and uniform initialization. Third row: Hessian iteration and kernel initialization (with $\mathbf{C} = \mathbf{K} + m_C^2\mathbf{I}$, $m_C^2 = 0.01$ and normalized afterwards for each value of x). Forth row: Gradient ($\mathbf{A} = \mathbf{I}$) with uniform initialization. Fifth row: Gradient with kernel initialization. Sixth row: Gradient with delta–peak initialization. (Initial L equal to $\ln(N/n + \epsilon)$, $\epsilon = 10^{-10}$, conditionally normalized. N/n is shown in Fig. 3.7.) Minimal number of iterations 4, maximal number of iterations 50, iteration stopped if $|L^{(i)} - L^{(i-1)}| < 10^{-8}$. Energy functional and parameters as for Fig. 3.5.

a misspecified prior and is typical for 'overfitting' or 'overtraining' (see, for example, [206,209]). For instance, in the fifth row of Fig. 3.9 the test error (and also the average training error) grows again after having reached a minimum while the energy is steadily decreasing.

3.5 Gaussian prior factors for general fields

3.5.1 The general case

In the previous sections we studied specific priors, Gaussian with respect to P or $L = \ln P$, and additional normalization (and non–negativity) conditions. In this section we consider the situation where the likelihood $p(y|x,\phi)$ is expressed in terms of some function $\phi(x,y)$. That means, we choose some (possibly non–linear) operator which maps the function ϕ to a likelihood, $P(x,y) = P(x,y;\phi)$. We then formulate the learning problem in terms of the hidden variables or unknown state of Nature ϕ.[2] We consider the case of a specific prior energy quadratic in ϕ

$$\frac{1}{2}\langle\phi|\mathbf{K}|\phi\rangle\,. \tag{3.98}$$

This means we are led to error functionals of the form

$$E(\phi) = -\langle \ln P(\phi)|\,N\rangle + \frac{1}{2}\langle\phi|\mathbf{K}|\phi\rangle + \langle P(\phi)|\,\Lambda_X\rangle\,, \tag{3.99}$$

where we have skipped the ϕ–independent part of the Λ_X–term. In the general case also the non–negativity constraint has to be included in Eq. (3.99), if it is active at the stationary solution and not already implemented by the chosen parameterization. Table 3.1 lists constraints to be implemented explicitly for some choices of ϕ.

To express the functional derivative of functional (3.99) with respect to ϕ we define besides the diagonal matrix $\mathbf{P} = \mathbf{P}(\phi)$ the Jacobi matrix of derivatives $\mathbf{P}' = \mathbf{P}'(\phi)$ with matrix elements

$$\mathbf{P}'(x,y;x',y';\phi) = \frac{\delta P(x',y';\phi)}{\delta\phi(x,y)}. \tag{3.100}$$

The matrix $\mathbf{P}'$ is diagonal for point–wise transformations where $p(y|x,\phi) = P(x,y;\phi) = P(\,\phi(x,y)\,)$. In such cases we can use P' to denote the vector of diagonal elements of $\mathbf{P}'$. An example is the previously discussed transformation $\phi = L = \ln P$, for which $P = e^{\phi}$ and $\mathbf{P}' = \mathbf{P}$. The stationarity equation for functional (3.99) becomes

$$0 = \mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N - \mathbf{K}\phi - \mathbf{P}'(\phi)\Lambda_X, \tag{3.101}$$

[2] Besides ϕ, the hyperparameters discussed in Chapter 5 also belong to the hidden variables.

ϕ	$P(\phi)$	constraints	
$P(x,y)$	$P = P$	norm	non–negativity
$z(x,y)$	$P = z/\int z\,dy$	—	non–negativity
$L(x,y) = \ln P$	$P = e^L$	norm	—
$g(x,y)$	$P = e^g/\int e^g\,dy$	—	—
$\phi = \int^y dy'\,P$	$P = d\phi/dy$	boundary	monotony

Table 3.1: Constraints for specific choices of ϕ.

and for existing $\mathbf{P}\mathbf{P}'^{-1} = (\mathbf{P}'\mathbf{P}^{-1})^{-1}$ (for nonexisting inverse see Section 4.1),

$$0 = N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi - \mathbf{P}\Lambda_X. \tag{3.102}$$

From Eq. (3.102) the Lagrange multiplier function can be found by integration, using the normalization condition $\mathbf{I}_X P = I$ in the form $\mathbf{I}_X\mathbf{P}\Lambda_X = \Lambda_X$ for $\Lambda_X(x) \neq 0$. Thus multiplying Eq. (3.102) by $\mathbf{I}_X$ yields

$$\Lambda_X = \mathbf{I}_X\left(N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi\right) = N_X - \mathbf{I}_X\mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi. \tag{3.103}$$

Λ_X is now eliminated by inserting Eq. (3.103) into Eq. (3.102)

$$0 = (\mathbf{I} - \mathbf{P}\mathbf{I}_X)\left(N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi\right). \tag{3.104}$$

Iteration procedures to solve Eq. (3.101),

$$\mathbf{K}\phi = T_\phi, \quad \text{with} \quad T_\phi(\phi) = \mathbf{P}'\mathbf{P}^{-1}N - \mathbf{P}'\Lambda_X, \tag{3.105}$$

are obtained by selecting a positive number η and a positive–definite matrix $\mathbf{A}$ (both may change during iteration) in

$$\phi^{(i+1)} = \phi^{(i)} + \eta\mathbf{A}^{-1}\left(T_\phi(\phi^{(i)}) - \mathbf{K}\phi^{(i)}\right). \tag{3.106}$$

Let us finally remark that the matrix notation we have used so far (and which we will continue to use) is not especially well–suited for calculating higher order derivatives for general ϕ. In that case higher order tensors appear which depend on more than two sets of coordinates, like the second derivative $\mathbf{P}''(x,y;x',y';x'',y'') = \delta^2 P(x'',y'')/(\delta\phi(x,y)\,\delta\phi(x',y'))$. For such

calculations it is convenient to let Greek indices denote the components in some orthonormal basis and to understand indices appearing in the same term as subscript and as superscript to be summed over, unless they appear only as multiple subscript or superscript, respectively. Then

$$E = -(\ln P)_\alpha N^\alpha + \frac{1}{2}\phi_\alpha \mathbf{K}^\alpha_\beta \phi^\beta + P_\alpha (\Lambda_X)^\alpha, \tag{3.107}$$

$$\delta_{\phi_\gamma} E = -(\mathbf{P}')^\gamma_\alpha (P^{-1})_\alpha N^\alpha + \mathbf{K}^\gamma_\alpha \phi^\alpha + (\mathbf{P}')^\gamma_\alpha (\Lambda_X)^\alpha, \tag{3.108}$$

$$(\Lambda_X)^\alpha = (\mathbf{I}_X)^\alpha_\beta N^\beta - (\mathbf{I}_X)^\alpha_\beta \mathbf{P}^\beta_\nu (\mathbf{P}'^{-1})^\nu_\mu \mathbf{K}^\mu_\tau \phi^\tau, \tag{3.109}$$

where $(\mathbf{P}')^\gamma_\alpha = \delta_{\phi_\gamma} P_\alpha$, and $(\mathbf{P}^{-1\prime})^\beta_\eta = \delta_{\phi^\eta}(P^{-1})^\beta = -(P^{-2})^\beta (\mathbf{P}')^\beta_\eta$. This notation enables us to write for the general Hessian for given Λ_X

$$\mathbf{H}^\gamma_\eta = -\delta_{\phi^\eta \phi_\gamma} E = -(\mathbf{P}'')^\gamma_{\beta\eta} \left[(P^{-1})^\beta N^\beta - (\Lambda_X)^\beta\right] - (\mathbf{P}')^\gamma_\beta (\mathbf{P}^{-1\prime})^\beta_\eta N^\beta + \mathbf{K}^\gamma_\eta, \tag{3.110}$$

with $(\mathbf{P}'')^\gamma_{\beta\eta} = \delta_{\phi^\eta}(\mathbf{P}')^\gamma_\beta$.

3.5.2 Square root of P

We already discussed in some detail Gaussian prior factors in log–likelihoods $\phi = \ln P$, where $P' = P = e^L$, $P/P' = 1$, and in likelihoods $\phi = P$, where $P' = 1$, $P/P' = P$. In the following we look at the choice $\phi = \sqrt{P}$, with

$$P = \phi^2, \quad P' = 2\phi, \quad P/P' = \phi/2, \tag{3.111}$$

which yields the common L_2–normalization condition over y

$$1 = \int dy\, \phi^2(x,y), \quad \forall x \in X, \tag{3.112}$$

being quadratic in ϕ. For real ϕ the non–negativity condition $P \geq 0$ is automatically satisfied [176,456].

For $\phi = \sqrt{P}$ and a negative Laplacian inverse covariance $\mathbf{K} = -\Delta$, one can relate the corresponding Gaussian prior to the *Fisher information* [13,15, 86,360,438,456]. Consider, for example, a problem with fixed x (so x can be omitted from the notation and one can write $P(y)$) and a d_y–dimensional y. Then one has, assuming the necessary differentiability conditions and vanishing boundary terms,

$$\langle \phi | \mathbf{K} | \phi \rangle = -\langle \phi | \Delta | \phi \rangle = \int dy \sum_k^{d_y} \left| \frac{\partial \phi}{\partial y_k} \right|^2 \tag{3.113}$$

$$= \sum_k^{d_y} \int \frac{dy}{4P(y)} \left(\frac{\partial P(y)}{\partial y_k} \right)^2 = \frac{1}{4} \sum_k^{d_y} I^F_k(0), \tag{3.114}$$

where $I_k^F(0)$ is the Fisher information, defined as

$$I_k^F(y_0) = \int dy \frac{\left|\frac{\partial P(y-y^0)}{\partial y_k^0}\right|^2}{P(y-y^0)} = \int dy \left|\frac{\partial \ln P(y-y^0)}{\partial y_k^0}\right|^2 P(y-y^0), \quad (3.115)$$

for the family $P(\cdot - y^0)$ with location parameter vector y^0.

A connection to quantum mechanics can be made for the case without training data where Eq. (3.99) becomes

$$E(\phi) = \frac{1}{2}\langle\phi|\mathbf{K}|\phi\rangle + \langle\Lambda_X|\phi^2\rangle, \quad (3.116)$$

for which the stationarity equation is homogeneous

$$\mathbf{K}\phi = -2\phi\Lambda_X. \quad (3.117)$$

For x–independent Λ_X this is an eigenvalue equation. Interpreting $\mathbf{K}$ as the Hamiltonian of a quantum mechanical system this becomes a Schrödinger equation with

$$-2\Lambda_X = \frac{\langle\phi|\mathbf{K}|\phi\rangle}{\langle\phi|\phi\rangle}, \quad (3.118)$$

representing the energy of state ϕ. In quantum mechanics Eq. (3.118) is the basis for variational methods (see also Section 4) to obtain approximate solutions for ground state energies [56, 121, 413].

Similar to the relation $\phi = \sqrt{P}$, the square ϕ^2 can be related to L (instead of P) by choosing $\phi = \sqrt{L_{\max} - L}$ for L bounded from above by $L_{\max}$. This corresponds to the normalization constraint

$$1 = \int dy\, e^{-\phi^2(x,y)+L_{\max}}, \quad \forall x \in X, \quad (3.119)$$

and yields $P = e^{-\phi^2+L_{\max}}$, $P' = -2\phi P$, $P/P' = -1/(2\phi)$.

3.5.3 Distribution functions

It is also interesting to formulate the prior in terms of the distribution function of the likelihood. The likelihood P for d_y–dimensional y–variable is obtained according to Eq. (3.78) from the distribution function ϕ by differentiation

$$P(\phi) = \prod_{k=1}^{d_y} \frac{\partial\phi}{\partial y_k} = \prod_{k=1}^{d_y} \nabla_{y_k}\phi = \prod_{k=1}^{d_y} \mathbf{R}_k^{-1}\phi. = \mathbf{R}^{-1}\phi, \quad (3.120)$$

resulting in a non–diagonal $\mathbf{P}'$. Eq. (3.120) corresponds to the transformation of (x–conditioned) density functions P in (x–conditioned) distribution functions $\phi = \mathbf{R}P$, or in components $\phi(x,y) = \int_{-\infty}^{y} P(x,y')dy'$. Because $\mathbf{R^T K R}$

is positive (semi–)definite if $\mathbf{K}$ is, a specific prior which is Gaussian in the distribution function ϕ is also Gaussian in the density P. $\mathbf{P}'$ becomes

$$\mathbf{P}'(x,y;x',y') = \frac{\delta\left(\prod_k^{d_y} \nabla_{y_k'} \phi(x',y')\right)}{\delta\phi(x,y)} = \delta(x-x') \prod_k^{d_y} \delta'(y_k - y_k'), \quad (3.121)$$

where the derivative of the δ–function is defined in Eq. (3.87). Fixing $\phi(x,-\infty) = 0$ the variational derivative $\delta/(\delta\phi(x,-\infty))$ is not needed. Thus working with the distribution function can be interesting because the constraints change their form. In particular, the normalization condition for P becomes for the distribution function $\phi = \mathbf{R}P$ the boundary condition $\phi(x,\infty) = 1$, $\forall x \in X$. The non–negativity condition for P corresponds to the monotonicity condition $\phi(x,y) \geq \phi(x,y')$, $\forall y \geq y'$, $\forall x \in X$ and to $\phi(x,-\infty) \geq 0$, $\forall x \in X$.

3.6 Covariances and invariances

3.6.1 Approximate invariance

Prior knowledge can often be related to approximate invariances, in particular to approximate symmetries. A Laplacian smoothness functional, for example, measures the deviation from translational symmetry under infinitesimal translations. Consider a linear mapping $\mathbf{T}$

$$\phi \to \tilde{\phi} = \mathbf{T}\phi. \quad (3.122)$$

To compare ϕ with $\tilde{\phi}$ we define a (semi–)distance by choosing a positive (semi–)definite $\mathbf{K}_T$, and use as error measure

$$\frac{1}{2} \langle \phi - \mathbf{T}\phi | \mathbf{K}_T | \phi - \mathbf{T}\phi \rangle = \frac{1}{2} \langle \phi | \mathbf{K} | \phi \rangle, \quad (3.123)$$

where

$$\mathbf{K} = (\mathbf{I} - \mathbf{T})^T \mathbf{K}_T (\mathbf{I} - \mathbf{T}). \quad (3.124)$$

$\mathbf{K}$ is positive semi–definite if $\mathbf{K}_T$ is. Conversely, every positive semi–definite $\mathbf{K}$ can be written $\mathbf{K} = \mathbf{W}^T\mathbf{W}$ and is thus of form (3.124) with $\mathbf{T} = \mathbf{I} - \mathbf{W}$ and $\mathbf{K}_T = \mathbf{I}$. Including terms of the form of (3.123) in a prior energy functional forces ϕ to be similar to $\tilde{\phi}$. We remind that for $\phi = P$ prior energies of the form (3.123) have been related in Section 2.3.2 (cf. Eq. (2.212) for $\mathbf{B} = \mathbf{W}^{-1}$) to input and output noise processes.

A special case are mappings leaving the norm invariant

$$\langle \phi | \phi \rangle = \langle \mathbf{T}\phi | \mathbf{T}\phi \rangle = \langle \phi | \mathbf{T}^T\mathbf{T}\phi \rangle. \quad (3.125)$$

For real ϕ and $\tilde{\phi}$, i.e., $(\mathbf{T}\phi) = (\mathbf{T}\phi)^*$, this requires $\mathbf{T}^T = \mathbf{T}^{-1}$ and $\mathbf{T}^* = \mathbf{T}$. In that case $\mathbf{T}$ has to be an orthogonal matrix and $\mathbf{K} = 2(\mathbf{I} - (\mathbf{T} + \mathbf{T}^T)/2)$

for $\mathbf{K}_T = \mathbf{I}$, where $(\mathbf{T} + \mathbf{T}^T)/2$ is the symmetric part of $\mathbf{T}$. An orthogonal matrix can be written as

$$\mathbf{T}(\theta) = e^{\mathbf{A}} = e^{\sum_i \theta_i \mathbf{A}_i} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\sum_i \theta_i \mathbf{A}_i \right)^k , \tag{3.126}$$

with antisymmetric $\mathbf{A} = -\mathbf{A}^T$ and real θ_i. Selecting a set of (generators) $\mathbf{A}_i$, Lie groups are obtained by varying the parameters θ_i in Eq. (3.126). Up to first order the expansion of the exponential function reads $\mathbf{S} \approx 1 + \sum_i \theta_i \mathbf{A}_i$. Thus we can define an error measure or specific prior energy with respect to infinitesimal transformations by

$$\begin{aligned} E_S = E(\phi|\tilde{D}_0) &= \frac{1}{2} \sum_{i,j} \left\langle \frac{\phi - (1 + \theta_i \mathbf{A}_i)\phi}{\theta_i} \middle| \mathbf{K}_{T,ij} \middle| \frac{\phi - (1 + \theta_j \mathbf{A}_j)\phi}{\theta_j} \right\rangle \\ &= \frac{1}{2} \left\langle \phi \middle| \sum_{i,j} \mathbf{A}_i^T \mathbf{K}_{T,ij} \mathbf{A}_j \middle| \phi \right\rangle , \end{aligned} \tag{3.127}$$

or, choosing for $\mathbf{K}_{T,ij} = \delta_{ij}\,\mathbf{I}$ the identity,

$$\begin{aligned} E_S = E(\phi|\tilde{D}_0) &= \frac{1}{2} \sum_{i} \left\langle \frac{\phi - (1 + \theta_i \mathbf{A}_i)\phi}{\theta_i} \middle| \frac{\phi - (1 + \theta_i \mathbf{A}_i)\phi}{\theta_i} \right\rangle \\ &= \frac{1}{2} \left\langle \phi \middle| \sum_{i} \mathbf{A}_i^T \mathbf{A}_i \middle| \phi \right\rangle . \end{aligned} \tag{3.128}$$

The prior energy (3.128) is typically multiplied by a factor λ when combined with training data terms. Summarizing, finite or infinitesimal approximate invariances can be implemented using prior energies of the form (3.123) or Eq. (3.127), respectively.

Let us consider the specific situation of an approximate invariance under coordinate transformations. In that case $\mathbf{T}$ represents a coordinate (or symmetry) transformation $\mathbf{S}$ which just changes the argument x of a function ϕ into $\sigma(x)$,

$$\mathbf{S}\phi(x) = \phi(\sigma(x)), \tag{3.129}$$

where $\sigma(x)$ is a permutation or one–to–one coordinate transformation. (Even for an arbitrary nonlinear coordinate transformation σ the corresponding operator $\mathbf{S}$ in the space of ϕ is linear.) An example is the translation of a function $\phi(x) \to \tilde{\phi}(x) = \mathbf{S}\phi(x) = \phi(x - c)$.

It is useful to see how $\mathbf{S}$ acts on the arguments of a function. Denoting the (improper) eigenvectors of the coordinate operator $\mathbf{x}$ with eigenvalue x by $\langle \cdot | x \rangle = |x\rangle$, i.e., $\mathbf{x}\,|x\rangle = x\,|x\rangle$, function values can be expressed as scalar products, e.g., $\phi(x) = \langle x|\,\phi\rangle$ for a function in x, or in two variables, $\phi(x, y)$

$= \langle x \otimes y | \phi \rangle$.[3] The action of $\mathbf{S}$ on a (real) function $\phi(x)$ is equivalent to the action of $\mathbf{S}^T$ ($= \mathbf{S}^{-1}$ if orthogonal) on the argument x

$$\mathbf{S}\phi(x) = \langle x | \mathbf{S}\phi \rangle = \langle \mathbf{S}^T x | \phi \rangle = \phi(\mathbf{S}^T x)^* = \phi(\mathbf{S}^T x), \tag{3.130}$$

or for $\phi(x, y)$

$$\mathbf{S}\phi(x, y) = \langle \mathbf{S}^T (x \otimes y) | \phi \rangle . \tag{3.131}$$

Writing $\mathbf{S} = \mathbf{S}_x \mathbf{S}_y$ we can split the action of $\mathbf{S}$,

$$\mathbf{S}\phi(x, y) = \langle (\mathbf{S}_x^T x) \otimes y | \mathbf{S}_y \phi \rangle . \tag{3.132}$$

This is convenient, for example, for vector fields in physics where x and $\phi(\cdot, y)$ form three–dimensional vectors with y representing a linear combination of component labels of ϕ. For the sake of simplicity, we consider in the following functions $\phi(x)$. Such a function can be a regression function, but other interpretations are possible. In the case of inverse quantum problems, for instance, $\phi(x)$ can represent a local potential $v(x)$. Considering y to be a component of a vector x, the results extend trivially also to functions $\phi(x, y)$.

Eq. (3.130) is also valid for general $\mathbf{T}$. For a general operator $\mathbf{T}$ the transformed argument $\mathbf{T}|x\rangle$ does not have to be an eigenvector of the coordinate operator $\mathbf{x}$ again. Hence $\mathbf{T}$ can map a specific $|x\rangle$ to arbitrary vectors being linear combinations of all $|x'\rangle$, i.e., $\mathbf{T}|x\rangle = \int dx' \, T(x, x') |x'\rangle$. A general orthogonal operator $\mathbf{T}$ maps an orthonormal basis to another orthonormal basis, but can still mix function values from different coordinates. Coordinate transformations, however, are represented by operators $\mathbf{S}$, which map coordinate eigenvectors $|x\rangle$ to other coordinate eigenvectors $|\sigma(x)\rangle$.

Special cases are linear coordinate transformations for which we can write $\phi(x) \to \tilde{\phi}(x) = \mathbf{S}\phi(x) = \phi(\mathbf{S}^T x) = \phi(S^T x)$, with a matrix S^T acting on a coordinate vector x (while the corresponding operator $\mathbf{S}^T$ acts in function space). Examples of such S^T are coordinate rotations $x' = S^T x$ like $x_1' = x_1 \cos\alpha - x_2 \sin\alpha$ and $x_2' = x_1 \sin\alpha + x_2 \cos\alpha$ for a two dimensional vector $x = (x_1, x_2)$. In analogy to Eq. (3.125) in the space of ϕ, they preserve the norm in coordinate space, $||x'||^2 = ||x||^2$, and in analogy to Eq. (3.126) coordinate rotations form a Lie group and we can write $\mathbf{S}(\theta) = e^{\sum_i \theta_i \mathbf{A}_i}$.

While exact symmetries play a fundamental role in physics, it will be shown in the following that approximate symmetries are a basic concept in empirical learning.

3.6.2 Infinitesimal translations

As a first example we show that a Laplacian smoothness prior can be interpreted as the implementation of an approximate symmetry under infinitesimal

[3] One may remark, that in this 'eigenvalue' notation, which is frequently used by physicists, $2|x\rangle \neq |2x\rangle$.

translations. Considering the group of translations in d dimensions, generated by the gradient operator ∇, this can be verified by recalling the multidimensional Taylor formula for expansion of a function $\phi(x)$ at (vector) x

$$\mathbf{S}(\theta)\phi(x) = e^{\sum_i \theta_i \nabla_i}\phi(x) = \sum_{k=0}^{\infty} \frac{(\sum_i \theta_i \nabla_i)^k}{k!}\phi(x) = \phi(x+\theta), \tag{3.133}$$

or analogously for a function $\phi(x, y)$ at x, y. Up to first order we have $\mathbf{S} \approx 1 + \sum_i \theta_i \nabla_i$. Hence for infinitesimal translations, the error measure or energy of Eq. (3.128) becomes

$$E_S = E(\phi|\tilde{D}_0) = \frac{1}{2}\sum_i \left\langle \frac{\phi - (1+\theta_i\nabla_i)\phi}{\theta_i} \middle| \frac{\phi - (1+\theta_i\nabla_i)\phi}{\theta_i} \right\rangle = -\frac{1}{2}\langle \phi | \Delta | \phi \rangle , \tag{3.134}$$

with Laplacian $\Delta = -\sum_i \nabla_i^T \nabla_i$, assuming vanishing boundary terms. This is the classical Laplacian smoothness term.

The discrete version of a negative Laplacian on a mesh with six points, with unit lattice spacing and periodic boundary conditions, has the structure

$$\mathbf{K} = -\Delta = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}. \tag{3.135}$$

The matrix (3.135) can be decomposed into square roots $\mathbf{K} = \mathbf{W}^T\mathbf{W}$ (see also Section 5.6) where a possible square root is

$$\mathbf{W} = \nabla_1^R = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}, \tag{3.136}$$

where ∇_1^R stands for the discrete difference operator defined as $\nabla_1^R\phi(x) = \phi(x+1) - \phi(x)$. In the continuum version the derivative operator $\partial/\partial x$ represents a square root of the negative Laplacian for periodic boundary conditions.

More generally, functions $\phi(x)$, which are smooth except for some discontinuities, can be approximated by eliminating matrix elements of the inverse covariance which connect the two sides of the discontinuity. (Another possibility is to use a Gaussian prior factor with discontinuous mean, see Section 3.7.) In inverse quantum problems such functions can represent potentials $\phi(x) = v(x)$ with expected discontinuities, corresponding, for example, to

possible dislocations on a smooth surface. Similarly, in natural images, edges of objects typically produce discontinuities in grey level intensities.

Two regions can be disconnected by deleting in (3.136) all lines of $\mathbf{W}$ which have matrix elements in both regions. For instance, the first three points in the six–dimensional space of Eq. (3.136) can be disconnected from the last three points by setting $\mathbf{W}(3,\cdot)$ and $\mathbf{W}(6,\cdot)$ to zero,

$$\tilde{\mathbf{W}} = \left(\begin{array}{cccccc} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right). \tag{3.137}$$

Squaring of $\tilde{\mathbf{W}}$ yields a positive semidefinite operator

$$\tilde{\mathbf{K}} = \tilde{\mathbf{W}}^T\tilde{\mathbf{W}} = \left(\begin{array}{cccccc} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{array}\right), \tag{3.138}$$

resulting in a smoothness prior which is ineffective between points from different regions. In contrast to using a prior with a discontinuous mean, the height of the jump at the discontinuity has not to be given in advance when working with disconnected Laplacians (or other disconnected inverse covariances). On the other hand training data are then required for all separated regions to determine the free constants which correspond to the zero modes of the local Laplacians. The reconstruction of discontinuous functions with non–Gaussian priors will be discussed in Section 6.

3.6.3 Approximate periodicity

As another example, lets us discuss the implementation of approximate periodicity. To measure the deviation from exact periodicity we define the difference operators

$$\nabla^R_\theta \phi(x) = \phi(x+\theta) - \phi(x), \tag{3.139}$$

$$\nabla^L_\theta \phi(x) = \phi(x) - \phi(x-\theta), \tag{3.140}$$

which are easily generalized to functions $\phi(x,y)$. For periodic boundary conditions $(\nabla^L_\theta)^T = -\nabla^R_\theta$, where $(\nabla^L_\theta)^T$ denotes the transpose of ∇^L_θ. Hence the operator

$$-\Delta_\theta = -\nabla^L_\theta\nabla^R_\theta = (\nabla^R_\theta)^T\nabla^R_\theta, \tag{3.141}$$

defined in analogy to the negative Laplacian, is positive (semi–)definite, and a possible error or energy term, enforcing approximate periodicity with period θ, is

$$\begin{aligned} E_S &= \frac{1}{2}\int dx\, |\phi(x) - \phi(x+\theta)|^2 \\ &= \frac{1}{2}\left\langle \nabla_\theta^R \phi \middle| \nabla_\theta^R \phi \right\rangle = -\frac{1}{2}\left\langle \phi \middle| \Delta_\theta \middle| \phi \right\rangle . \end{aligned} \tag{3.142}$$

It is instructive to display Δ_θ for discretized ϕ, where the operator ∇_θ^R for periodic boundary conditions becomes, for the example on a mesh with six points and $\theta = 2$, the matrix

$$\nabla_\theta^R = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}, \tag{3.143}$$

so that

$$-\Delta_\theta = \begin{pmatrix} 2 & 0 & -1 & 0 & -1 & 0 \\ 0 & 2 & 0 & -1 & 0 & -1 \\ -1 & 0 & 2 & 0 & -1 & 0 \\ 0 & -1 & 0 & 2 & 0 & -1 \\ -1 & 0 & -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & -1 & 0 & 2 \end{pmatrix}. \tag{3.144}$$

As every periodic function with $\phi(x) = \phi(x+\theta)$ is in the null space of Δ_θ, besides the data terms another error term has to be added to get a unique solution of the stationarity equation. Choosing, for example, a Laplacian smoothness term, yields specific prior energies

$$E_S = -\frac{\lambda}{2}\left\langle \phi \middle| \gamma\Delta + (1-\gamma)\Delta_\theta \middle| \phi \right\rangle , \tag{3.145}$$

with inverse covariance $\mathbf{K} = -\lambda(\gamma\Delta + (1-\gamma)\Delta_\theta)$ and weighting factors $\lambda > 0$, $0 \le \gamma \le 1$. In case θ, λ, γ are not known, they can be treated as hyperparameters as discussed in Section 5. It is also possible to sum over several periods

$$E_S = \frac{1}{2}\sum_{k=1}^{k_{\max}} w(k) \int dx\, |\phi(x) - \phi(x+k\theta)|^2, \tag{3.146}$$

where $w(k)$ is a weight function, decreasing for larger k. Prior energies as in (3.146) enforce approximate periodicity over longer distances than a prior energy of the form (3.142). The latter, on the other hand, is more robust than

(3.146) with respect to local deviations from periodicity, like a locally varying frequency. An example of the reconstruction of an approximately periodic regression function is given in Section 3.8.5.

Alternatively, approximate symmetries can be implemented by giving explicitly a symmetric reference or template function t in specific prior energies like $\frac{1}{2}\langle\phi - t|\mathbf{K}|\phi - t\rangle$. In case we are expecting an approximate periodic function $\phi(x)$ this would be a template function with $t(x) = t(x + \theta)$. This possibility will be discussed in Section 3.7.

3.6.4 Approximate fractals

Nice looking fractal images can be produced by Multiple Reduction Copy Machines (MRCM) being examples of so–called Iterated Function Systems [327, 391]. Each iteration step of an MRCM consists of placing M reduced affine copies of the whole image on specific regions of the original image. A fractal image is then produced by iterating this multiple affine copy operation.

For instance, consider a one–dimensional function defined on an interval L and a multiple reduced copying operation which places M down–scaled copies of the whole interval L in M (non–overlapping) subintervals $A_k = [a_k, b_k]$, thus reducing the interval L by scaling factors $|A_k|/|L| = s_k < 1$. Such an operation can be written as

$$\mathbf{C} = (\mathbf{I} - \sum_{k=1}^{M} \pi_k) + \sum_{k=1}^{M} \mathbf{C}_k \pi_k, \tag{3.147}$$

where π_k is the projector on region A_k and each $\mathbf{C}_k$ represents a reduced copy operation of the interval L into one of the A_k, i.e.,

$$\phi \to \mathbf{C}_k\phi(x) = \phi(\frac{x - a_k}{s_k}). \tag{3.148}$$

Using the formulation of Eq. (3.147) we have chosen to leave the function unchanged in regions not contained in any of the sets A_k. Approximate invariance under m repeated applications of the multiple reduced copy operation $\mathbf{C}$ is, according to Eq. (3.124), implemented by the inverse covariance

$$\mathbf{K}_{\mathrm{F1}}^{(m)} = (\mathbf{I} - \mathbf{C}^m)^T(\mathbf{I} - \mathbf{C}^m). \tag{3.149}$$

To create a perfect fractal the operation $\mathbf{C}$ has to be iterated an infinite number of times. Clearly, on a discrete lattice a reduced copy operation can only be performed a finite number of times. In particular, down scaling a function on a finite lattice means loss of information (the discrete version of the transformation $\mathbf{C}$ is not invertible). An effective operation $\mathbf{C}$ can be implemented in the form

$$\mathbf{C} = (\mathbf{I} - \sum_{k=1}^{M} \pi_k) + \sum_{k=1}^{M} \mathbf{C}_k \mathbf{A}_k \pi_k, \tag{3.150}$$

where $\mathbf{A}_k$ stands for an averaging operation $\mathbf{A}_k\phi(x) = \sum_{x'} K_k(x-x')\phi(x')$ with some real, positive and normalized kernel K_k. Taking integer $1/s_k$, we can, for example, uniformly average over all the $1/s_k$ points of ϕ which contribute to the same point of the down–scaled function ϕ'

$$\mathbf{A}_k\phi(x) = s_k \sum_{x'=x}^{x-1+1/s_k} \phi(x'). \tag{3.151}$$

Due to the averaging in Eq. (3.150) the classical one–dimensional Cantor set (considered as binary function of x), for instance, is not invariant under $\mathbf{C}^m$ on a discrete lattice. It is however invariant under the following operation

$$\mathbf{K}_{\mathrm{F2}}^{(m)} = \sum_{k=1}^{M^m} (\mathbf{I} - \mathbf{C}^m\mathbf{B}_k)^T(\mathbf{I} - \mathbf{C}^m\mathbf{B}_k), \tag{3.152}$$

where the sum is over all M^m (nonoverlapping) subintervals of the mth iteration level and $\mathbf{B}_k$ denotes the 'blowing up' operation which maps the subinterval A_k to the whole interval L

$$\phi'(x) = \mathbf{B}_k\,\phi(x) = \phi(s_k x + a_k). \tag{3.153}$$

In a discrete implementation of (3.153) the function $\phi(s_k x + a_k)$ for noninteger $x' = s_k x + a_k$ has either to be interpolated or replaced by its value at, e.g., the nearest or next lower integer. An inverse covariance of the form (3.152) implements the expectation to find the same small scale structures in all M^m subintervals A_k. To allow a stronger influence of lower iteration levels, inverse covariances of the form (3.152) can also be added, e.g., for all levels $1 \le l \le m$,

$$\mathbf{K}_{\mathrm{F3}}^{(m)} = \sum_{l=1}^{m} \mathbf{K}_{\mathrm{F2}}^{(l)} = \sum_{l=1}^{m} \sum_{k=1}^{M^l} (\mathbf{I} - \mathbf{C}^l\mathbf{B}_k)^T(\mathbf{I} - \mathbf{C}^l\mathbf{B}_k). \tag{3.154}$$

An example of the reconstruction of an approximately fractal regression function can be found in Section 3.8.5. As a fractal (or multifractal) structure has been proposed for financial time series, this may be an interesting application area for approximate fractal symmetries [328].

3.7 Non–zero means

A prior energy term $\langle\phi\,|\mathbf{K}|\,\phi\rangle\,/2$ measures the squared $\mathbf{K}$–distance of ϕ to the zero function $t \equiv 0$. Choosing a zero mean function for the prior process is calculationally convenient for Gaussian priors, but by no means mandatory. In particular, a function ϕ is in practice often measured relative to some non–trivial base line. Without further *a priori* information that base line can in

principle be an arbitrary function. Choosing a zero mean function, that base line does not enter the formulae and appears only in the realization of the measurement process. On the other hand, including explicitly a non–zero mean function t, playing the role of a function *template* (or reference, target, prototype, base line) can be a very powerful tool. It allows, for example, to parameterize $t(x, y; \theta)$ by introducing hyperparameters θ (see Section 5) and to specify explicitly different maxima of multimodal priors (see Section 6 and [291–295].) All that can not be done by referring to a single baseline function $t \equiv 0$.

Hence in this section we consider energy/error terms of the form

$$\frac{1}{2} \langle \phi - t \,|\mathbf{K}|\, \phi - t \rangle \tag{3.155}$$

corresponding to a Gaussian prior factor

$$p(\phi|D_0) = \left(\det \frac{\mathbf{K}}{2\pi} \right)^{\frac{1}{2}} e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t \rangle}. \tag{3.156}$$

Mean (reference, template) functions t allow an easy and straightforward implementation of prior information in form of examples for ϕ. They are the continuous analogue of standard training data. The fact that template functions t are often chosen equal to zero, and thus do not appear explicitly in the error functional, should not obscure the fact that they are of key importance for any generalization. There are many situations where it can be very valuable to include non–zero prior means explicitly. For example, a natural candidate for a template function $t(x, y)$, or $t(x)$ for regression, is the output of an empirical learning system (like a neural net, decision tree, or nearest neighbor method) resulting from learning the same or a similar task in the past. In this way template functions can be used for *transfer* of knowledge between similar problems or to *include the results of earlier learning* on the same problem, e.g., in case the original data are lost but the output of another learning system is still available. In particular, the parameters θ of a parametric approach, for example, the weights of a neural network, can be used as hyperparameters. As shown in Section 5.3 this allows the adaption of parametric template functions $t(x, y; \theta)$ (or $t(x; \theta)$ for regression problems) using the available training data D_T. A detailed discussion of possibilities to select template functions can be found in [294]. (See also [291–293, 295].)

Including non–zero template functions generalizes functional $E(\phi)$ of (3.99) to

$$\begin{aligned} E(\phi) &= -\langle \ln P(\phi) |\, N \rangle + \frac{1}{2} \langle \phi - t \,|\mathbf{K}|\, \phi - t \rangle + \langle P(\phi) |\, \Lambda_X \rangle \qquad (3.157) \\ &= -\langle \ln P(\phi) |\, N \rangle + \frac{1}{2} \langle \phi \,|\mathbf{K}|\, \phi \rangle - \langle J |\, \phi \rangle + \langle P(\phi) |\, \Lambda_X \rangle + c. \end{aligned}$$

In the language of physics $J = \mathbf{K}t$ represents an *external field* coupling to $\phi(x, y)$, similar, for example, to a magnetic field. A non–zero field J leads to a non–zero expectation of ϕ in the no–data case. (The ϕ–independent c stands for the term $\frac{1}{2}\langle t|\mathbf{K}|t\rangle$, which is for invertible $\mathbf{K}$ equal to $\frac{1}{2}\langle J|\mathbf{K}^{-1}|J\rangle$, and can be skipped when minimizing $E(\phi)$ with respect to ϕ.) The stationarity equation for $E(\phi)$ with non–zero template t contains now an inhomogeneous term $\mathbf{K}t = J$

$$0 = \mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N - \mathbf{P}'(\phi)\Lambda_X - \mathbf{K}(\phi - t), \tag{3.158}$$

with

$$\Lambda_X = \mathbf{I}_X\left(N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}(\phi - t)\right), \tag{3.159}$$

for invertible $\mathbf{P}\mathbf{P}'^{-1}$ and $\Lambda_X \neq 0$. Note that functional (3.157) can be rewritten as a functional with zero template $t \equiv 0$ in terms of $\widetilde{\phi} = \phi - t$. This substitution, however, changes the likelihood terms and constraints. We will see in Section 6 that, when using non–additive combinations of squared distances of the form (3.155) with different non–zero templates, those can not be removed by such a substitution. Additive combinations of squared error terms, on the other hand, can again be written as one squared error term, using a generalized 'bias–variance' decomposition

$$\frac{1}{2}\sum_{j=1}^{n}\langle\phi - t_j|\mathbf{K}_j|\phi - t_j\rangle = \frac{1}{2}\langle\phi - t|\mathbf{K}|\phi - t\rangle + E_{\min} \tag{3.160}$$

with *template average*

$$t = \mathbf{K}^{-1}\sum_{j=1}^{n}\mathbf{K}_j t_j, \tag{3.161}$$

assuming the existence of the inverse of the operator

$$\mathbf{K} = \sum_{j=1}^{n}\mathbf{K}_j, \tag{3.162}$$

and *minimal energy/error*

$$E_{\min} = \frac{n}{2}V(t_1, \cdots t_N) = \frac{1}{2}\left(\sum_{j=1}^{n}\langle t_j|\mathbf{K}_j|t_j\rangle - \langle t|\mathbf{K}|t\rangle\right), \tag{3.163}$$

which up to a factor $n/2$ represents a generalized template variance V. We end with the remark that adding error terms corresponds in its probabilistic Bayesian interpretation to ANDing independent events. For example, if we wish to implement that ϕ is likely to be smooth AND mirror symmetric, we may add two squared error terms, one related to smoothness and another to

mirror symmetry. According to (3.160) the result will be a single squared error term of form (3.155).

As mentioned in Section 3.6.3 approximate symmetries can also be implemented by using a symmetric template or reference function $t = \mathbf{S}t$ in a Gaussian prior factor instead of a specifically designed covariance. For regression problems with approximate periodicity, this means choosing a periodic template function $t(x) = t(x+\theta)$ in the prior energy $E_S = \frac{1}{2}\langle \phi - t | \mathbf{K} | \phi - t \rangle$ where $\mathbf{K}$ could be the identity or a differential operator. (For density estimation problems approximate periodicity can be formulated with respect to x and y and be implemented by using a corresponding template $t(x,y)$.) Thus a periodic reference potential favors a specific form for the reconstructed potential, including a specific frequency and phase. This is different for the covariance implementation (3.142) of approximate periodicity where only the frequency is relevant and reference potentials can still be chosen arbitrarily. They may as well be nonperiodic functions or functions with even higher symmetry like $t \equiv 0$ which is invariant under all translations. Flexible reference functions will be studied in Section 5.

3.8 Regression

3.8.1 Gaussian regression

A special case of density estimation is regression where the likelihood is completely determined by its regression function. In particular, we speak of *Gaussian regression* if the likelihoods are Gaussian so that for independent training data and one–dimensional y (and possibly multidimensional x)

$$p(y_i|x_i,\phi) = \frac{1}{\sqrt{2\pi}\sigma(x_i)} e^{-\frac{|y_i - \phi(x_i)|^2}{2\sigma^2(x_i)}} \propto e^{-E_i}, \tag{3.164}$$

with known variance σ^2, which may be independent of x as in Eq. (2.11), but can also be chosen dependent on x, and single trial training energy

$$E_i = \frac{|y_i - \phi(x_i)|^2}{2\sigma^2(x_i)}. \tag{3.165}$$

Analogously, for d–dimensional y

$$\begin{aligned} p(y_i|x_i,\phi) &= (2\pi)^{-\frac{d}{2}} (\det C(x_i))^{-\frac{1}{2}} e^{-\frac{1}{2}\sum_{k,l=1}^{d} (\phi_k(x_i) - y_{ik}) C^{-1}_{kl}(x_i) (\phi_l(x_i) - y_{il})} \\ &= (2\pi)^{-\frac{d}{2}} (\det C(x_i))^{-\frac{1}{2}} e^{-\frac{1}{2}||y_i - \phi(x_i)||^2_{C^{-1}(x_i)}} \propto e^{-E_i}, \end{aligned} \tag{3.166}$$

with a positive definite $(d \times d)$ sampling covariance matrix $C(x)$ and single trial training energy

$$E_i = \frac{1}{2}||y_i - \phi(x_i)||^2_{C^{-1}(x_i)} = \frac{1}{2}\sum_{k,l=1}^{d}(\phi_k(x_i) - y_{ik})C^{-1}_{kl}(x_i)(\phi_l(x_i) - y_{il}). \tag{3.167}$$

For independent components, with a sampling variance which depends on x and on the dimension k, for example, $C_{kl}(x) = \delta_{kl}\sigma_k^2(x)$. (In general, one can also work with sampling covariances $C_{kl,i}(x)$ which depend on the trial i which allows varying covariances for the same x. For predictions then the covariances for the test data also have to be known.)

Similarly, in classical inverse problems [122, 221, 262, 478, 492, 494] one is interested in the reconstruction of an underlying function ϕ which is related to an observable $\tilde{\phi} = \mathbf{F}\phi$ with a typically ill–conditioned (or even singular) $\mathbf{F}$. In that case one considers single trial training energies of the form

$$E_i = \frac{1}{2\sigma^2(x_i)}\left|\int dx'\,\mathbf{F}(x_i, x')\phi(x') - y_i\right|^2, \tag{3.168}$$

where in the general multidimensional case $|\cdot|^2/(2\sigma^2)$ has to be replaced by $||\cdot||^2_{C^{-1}(x)}$. For example, the linear operator $\mathbf{F}$ can represent an averaging process which models the 'instrument or apparatus function' of a measuring device. Examples of such linear inverse problems include image restoration, X–ray computer tomography, and Fourier synthesis in radio astronomy (see the above references and [348]). In that case the regression function of the observed values is $\mathbf{F}\phi$, while $\phi(x)$ represents the underlying true function one is interested in. For some problems one also has to consider nonlinear mappings $\mathbf{F}(\phi)$, which destroy the quadratic structure of the reconstruction problem and typically all calculations have to be done numerically. Similar to C also $\mathbf{F}$ can be chosen i–dependent.

In (3.168) the observed data are nonlocal in the function $\phi(x)$, meaning that a specific y_i does not depend on one specific value $\phi(x_i)$ only, but possibly on the whole function $\phi(x)$. Nonlocal dependencies like smoothness are thus implemented by the data term, i.e., by a specific choice of $\mathbf{F}$. It is then possible to use a simple maximum entropy prior of the form (2.111) for ϕ and still get a smooth ϕ.

Clearly there are problems where it is more appropriate to work with non–Gaussian likelihoods. If the form of the likelihood is fixed except for an unknown mean function we speak of a *non–Gaussian regression problem.* For instance, in the analysis of empirical spectra a Poisson distribution is often a better model for the likelihood than a Gaussian. Reconstructing the mean function $\phi(x) > 0$ of Poisson–distributed likelihoods (2.13),

$$p(y|x, \phi) = [\phi(x)]^y e^{-\phi(x)}/y!,\ y \in N \tag{3.169}$$

in the analysis of spectral data (with counting rates represented by integer y) is an example of a non–Gaussian regression problem [65,314,457].

Because in Gaussian regression the training energies (3.164), (3.166), or (3.168) are like the prior energies quadratic in ϕ, it is convenient to write training and prior energies in an analogous form. To this end, we define for training data (x_i, y_i) the training data templates

$$t_i \equiv y_i, \tag{3.170}$$

and the operator

$$\mathbf{K}_i(x, x') = C^{-1}\delta(x - x_i)\delta(x - x'), \tag{3.171}$$

where for the $\mathbf{K}_i$ corresponding to (3.164) $C = \sigma^2$. The training energies in (3.164) and (3.166) can then be written, in a form analogous to a quadratic prior energy, as

$$E_i = \frac{1}{2}\langle \phi - t_i \,|\mathbf{K}_i|\, \phi - t_i\rangle_X \,. \tag{3.172}$$

The subscript X indicating that only x–integration is required to calculate the matrix elements will, for the sake of simplicity, be skipped in the following. (We have already remarked in Section 2.1.6 that, for continuous x, measurement for $\phi(x)$ has to be understood as measurement for $\phi(\tilde{x}) = \int dx\, \vartheta(x)\phi(x)$ for sharply peaked $\vartheta(x)$. We assume that the discretization of ϕ used in numerical calculations takes care of that averaging.) Analogous to Eq. (3.168) we obtain, with possibly multidimensional y,

$$E_i = \frac{1}{2}\langle \mathbf{F}\phi - t_i \,|\mathbf{K}_i|\, \mathbf{F}\phi - t_i\rangle \,. \tag{3.173}$$

As in Eq. (3.160) the complete training energy E_D for n data points, sampled according to (3.164) or (3.166), becomes

$$\begin{aligned} E_D = \frac{1}{2}\chi^2 &= \frac{1}{2}\sum_{i=1}^{n} ||\phi(x_i) - y_i||_{C^{-1}(x_i)} \\ &= \frac{1}{2}\langle \phi - t_D \,|\mathbf{K}_D|\, \phi - t_D\rangle + E^D_{\min}, \end{aligned} \tag{3.174}$$

(χ^2 corresponds to the usual 'chi–squared' of least squares minimization [410]), with inverse data covariance

$$\mathbf{K}_D = \sum_{i=1}^{n} \mathbf{K}_i, \tag{3.175}$$

data template (averaged data vector)

$$t_D = \mathbf{K}_D^{-1} \sum_{i=1}^{n} \mathbf{K}_i t_i, \tag{3.176}$$

(weighted) data vector

$$\mathbf{K}_D\, t_D = \sum_{i=1}^{n} \mathbf{K}_i\, t_i, \tag{3.177}$$

and ϕ–independent minimal error

$$E^D_{\min} = \frac{1}{2}\left(\sum_{i=1}^{n} \langle t_i\,|\mathbf{K}_i|\,t_i\rangle - \langle t_D\,|\mathbf{K}_D|\,t_D\rangle\right), \tag{3.178}$$

being proportional to the sum of the local variances of the normalized data

$$V_D = 2E^D_{\min} = \sum_{i}^{n} V_i, \tag{3.179}$$

where for (3.164)

$$V_i = \sigma^{-2}(x_i)\left(\frac{1}{n_i}\sum_{k=1}^{n_i} y_k(x_i)^2 - \left(\frac{1}{n_i}\sum_{k=1}^{n_i} y_k(x_i)\right)^2\right), \tag{3.180}$$

is the normalized variance of the n_i training data at x_i. If all x_i are different then $V_D = 0$. Indeed, for (3.164) the inverse data covariance is

$$\mathbf{K}_D(x,x') = \frac{1}{\sigma^2(x)}\sum_{i=1}^{n}\delta(x-x_i)\delta(x-x') = \frac{n_x}{\sigma^2(x)}\delta(x-x'), \tag{3.181}$$

where $n_x = N_X(x) = \sum_{i=1}^{n}\delta(x-x_i)$ is zero for x not in the training data and otherwise (for discrete x where the δ–function is understood as Kronecker–δ or for continuous x after integration over a sufficiently peaked test function) the number of measurements at x. The expression $\mathbf{K}_D^{-1} = (\mathbf{I}_D\mathbf{K}_D\mathbf{I}_D)^{-1} = \sigma^2(x)\delta(x-x')/n_x$ in (3.176) denotes the inverse within the space of training data. Here we used the projector $\mathbf{I}_D$ into the space of training data, introduced in Section 3.1.2,

$$\mathbf{I}_D(x,x') = \delta(x-x')\sum_{i=1}^{\tilde{n}}\delta(x-x_i). \tag{3.182}$$

where the sum is over all $\tilde{n} \leq n$ different training data x_i. Thus number $\tilde{n}$ is smaller than the number n of all training data if two or more x_i are equal, and $\mathbf{K}_D = \mathbf{I}_D$ for $n = \tilde{n}$ and $\sigma(x) = 1$. (Again, for continuous x an additional integration of x_i over a small interval I_i is required to make $\mathbf{I}_D^2$ well defined.) The data template t_D for (3.164) becomes the mean of y–values measured at x

$$t_D(x) = \frac{1}{n_x}\sum_{\substack{j=1\\ x_j=x}}^{n_x} y(x_j), \tag{3.183}$$

and $t_D(x) = 0$ for $n_x = 0$. Similar to Eq. (3.174), we obtain for (3.168)

$$E_D = \frac{1}{2} \langle \mathbf{F}\phi - t_D \,|\mathbf{K}_D|\, \mathbf{F}\phi - t_D \rangle + E^D_{\min}. \tag{3.184}$$

For existing inverse $\mathbf{F}^{-1}$, the ϕ–dependent term can also be written in the form

$$E_D = \frac{1}{2} \langle \phi - \mathbf{F}^{-1}t_D \,|\mathbf{F}^T\mathbf{K}_D\mathbf{F}|\, \phi - \mathbf{F}^{-1}t_D \rangle \tag{3.185}$$

which takes its minimum with respect to ϕ at $\phi = \mathbf{F}^{-1}t_D$.

In Gaussian regression the logarithmic data term $\sum_i \ln P(x_i, y_i)$ becomes quadratic in the regression function $\phi(x_i)$. Hence, if data terms of the form (3.164) or (3.166) are combined with the prior energy of a Gaussian prior factor, then both terms have the quadratic form (3.155). In that case the posterior is Gaussian in ϕ, the posterior energy is quadratic in ϕ, and the stationarity equation is linear. This situation is very convenient for actual calculation.

We now combine Gaussian data terms and a Gaussian (specific) prior $p(\phi|D_0)$, in short $p(\phi)$, with prior operator $\mathbf{K}_0(x, x')$. If we allow a non–zero prior template t_0 the resulting energy functional can, according to (3.160), be written in the following forms, useful for various purposes,

$$\begin{aligned}
E(\phi) &= \frac{1}{2}\sum_{i=1}^{n} ||\phi(x_i) - y_i||^2_{C^{-1}(x_i)} + \frac{1}{2}\langle \phi - t_0 \,|\mathbf{K}_0|\, \phi - t_0 \rangle && (3.186)\\
&= \frac{1}{2}\sum_{i=1}^{n} \langle \phi - t_i \,|\mathbf{K}_i|\, \phi - t_i \rangle + \frac{1}{2}\langle \phi - t_0 \,|\mathbf{K}_0|\, \phi - t_0 \rangle && (3.187)\\
&= \frac{1}{2}\sum_{i=0}^{n} \langle \phi - t_i \,|\mathbf{K}_i|\, \phi - t_i \rangle && (3.188)\\
&= \frac{1}{2}\langle \phi - t_D \,|\mathbf{K}_D|\, \phi - t_D \rangle + \frac{1}{2}\langle \phi - t_0 \,|\mathbf{K}_0|\, \phi - t_0 \rangle + E^D_{\min} && (3.189)\\
&= \frac{1}{2}\langle \phi \,|\mathbf{K}|\, \phi \rangle - \langle \phi \,|\mathbf{K}|\, t \rangle + \frac{1}{2}\sum_{i=0}^{n} \langle t_i \,|\mathbf{K}_i|\, t_i \rangle && (3.190)\\
&= \frac{1}{2}\langle \phi - t \,|\mathbf{K}|\, \phi - t \rangle + E_{\min}, && (3.191)
\end{aligned}$$

with total inverse covariance

$$\mathbf{K} = \sum_{i=0}^{n} \mathbf{K}_i, \tag{3.192}$$

total template

$$t = \mathbf{K}^{-1}\sum_{i=0}^{n} \mathbf{K}_i t_i, \tag{3.193}$$

total weighted data vector

$$\mathbf{K}t = \sum_{i=0}^{n} \mathbf{K}_i t_i, \tag{3.194}$$

and ϕ–independent minimal error

$$E_{\min} = \frac{1}{2}\left(\sum_{i=0}^{n} \langle t_i \,|\mathbf{K}_i|\, t_i\rangle - \langle t\,|\mathbf{K}|\,t\rangle\right), \tag{3.195}$$

being proportional to the 'generalized variance' $V = 2E_{\min}/(n+1)$. In all matrix elements or scalar products integration is over x–variables only. Normalization of $P(x,y)$ is not influenced by a change in $\phi(x)$, so the Lagrange multiplier terms have been skipped.[4]

The stationarity equation is most easily obtained from (3.191),

$$0 = \mathbf{K}(\phi - t). \tag{3.196}$$

It is linear and has, on a space where $\mathbf{K}^{-1}$ exists, the unique solution

$$\phi = t. \tag{3.197}$$

From the definition (3.193) of t we see that the MAP solution is not changed if the data variance σ^2 or covariance C is absorbed in the prior covariance, i.e., under the transformations $\mathbf{K}_D \to \sigma^2\mathbf{K}_D$, $\mathbf{K}_0 \to \sigma^2\mathbf{K}_0$, or $\mathbf{K}_D \to C\mathbf{K}_D$, $\mathbf{K}_0 \to C\mathbf{K}_0$. Thus we can choose $\sigma = 1$ or $C = 1$, respectively, without restricting generality.

Analogously, for a posterior energy of the form

$$E(\phi) = \frac{1}{2}\langle \mathbf{F}_D\phi - t_D \,|\mathbf{K}_D|\, \mathbf{F}_D\phi - t_D\rangle + \frac{1}{2}\langle \mathbf{F}_0\phi - t_0 \,|\mathbf{K}_0|\, \mathbf{F}_0\phi - t_0\rangle \tag{3.198}$$

(which is similar in structure to the energy of the linear regression model (4.74) discussed in Section 4.4), one obtains as MAP equation

$$\mathbf{F}_D^T\mathbf{K}_D(\mathbf{F}_D\phi - t_D) = -\mathbf{F}_0^T\mathbf{K}_0(\mathbf{F}_0\phi - t_0) \tag{3.199}$$

with MAP solution $\phi = t$ where now

$$t = \mathbf{K}^{-1}\left(\mathbf{F}_D^T\mathbf{K}_D t_D + \mathbf{F}_0^T\mathbf{K}_0 t_0\right), \tag{3.200}$$

with

$$\mathbf{K} = \mathbf{F}_D^T\mathbf{K}_D\mathbf{F}_D + \mathbf{F}_0^T\mathbf{K}_0\mathbf{F}_0. \tag{3.201}$$

Data y_i are in this case approximated by

$$\tilde{\phi}(x_i) = \langle x_i \,|\mathbf{F}_D|\, \phi\rangle = \int dx\, \mathbf{F}_D(x_i, x)\,\phi(x), \tag{3.202}$$

[4]We remind the reader that in regression we use the bra–ket notation for matrix elements and scalar products to refer to integration over x only.

for which we find

$$\begin{aligned}\widetilde{\phi} = \mathbf{F}_D\,\phi &= \mathbf{F}_D\mathbf{K}^{-1}\left(\mathbf{F}_D^T\mathbf{K}_D t_D + \mathbf{F}_0^T\mathbf{K}_0 t_0\right)\\ &= \mathbf{A}\,\mathbf{K}_D\,t_D \;+\; \mathbf{B}\,\mathbf{K}_0\,t_0,\end{aligned} \tag{3.203}$$

defining

$$\mathbf{A} = \mathbf{F}_D\mathbf{K}^{-1}\mathbf{F}_D^T = \mathbf{F}_D\left(\mathbf{F}_D^T\mathbf{K}_D\mathbf{F}_D + \mathbf{F}_0^T\mathbf{K}_0\mathbf{F}_0\right)^{-1}\mathbf{F}_D^T, \tag{3.204}$$

$$\mathbf{B} = \mathbf{F}_D\mathbf{K}^{-1}\mathbf{F}_0^T = \mathbf{F}_D\left(\mathbf{F}_D^T\mathbf{K}_D\mathbf{F}_D + \mathbf{F}_0^T\mathbf{K}_0\mathbf{F}_0\right)^{-1}\mathbf{F}_0^T. \tag{3.205}$$

In the case of a zero template $t_0 = 0$ the operator $\mathbf{A}$ maps the (weighted) data vector $\mathbf{K}_D t_D$ to the approximation $\widetilde{\phi}$

$$\widetilde{\phi} = \mathbf{A}\mathbf{K}_D\,t_D. \tag{3.206}$$

Referring to kernel smoothing [162, 198, 203, 455], $\mathbf{A}(x, x')$ may be called an equivalent kernel or, its restriction to data space $\mathbf{A}_{DD} = \mathbf{I}_D\mathbf{A}\mathbf{I}_D$, influence matrix [507]. Analogously, $\mathbf{B}$ may be called 'prior template influence operator', and $\mathbf{F}_D\mathbf{K}^{-1}$ 'total influence operator' with respect to $\sum_{i=0}^{n}\mathbf{K}_i t_i$ because $\widetilde{\phi} = \mathbf{F}_D\mathbf{K}^{-1}\left(\sum_{i=0}^{n}\mathbf{K}_i t_i\right)$.

The operator $\mathbf{K}$ can be invertible (and usually is, hence the MAP solution is unique) even if $\mathbf{K}_0$ is not invertible. The inverse $\mathbf{K}^{-1}$, necessary to calculate t, is training data dependent and represents the covariance operator/matrix of a Gaussian posterior process. In many practical cases, however, the prior covariance $\mathbf{K}_0^{-1}$ (or in case of a null space a pseudo inverse of $\mathbf{K}_0$) is directly given or can be calculated. Then, as shown in the following, an inversion of a finite dimensional matrix in data space is sufficient to find the minimum of the energy $E(\phi)$ [162, 507]. Direct inversion of a $\tilde{n}^3 \times \tilde{n}^3$–matrix scales with $O(\tilde{n}^3)$ (one can do a little bit better), iterative inversion methods can scale with $O(\tilde{n}^2)$ [154, 170, 410]. (Lower dimensional approximations for Gaussian processes, useful for large $\tilde{n}$, are discussed in [486, 536].)

Invertible $\mathbf{K}_0$: Let us first study the case of an invertible inverse prior covariance $\mathbf{K}_0$ and consider the stationarity equation as obtained from (3.187) or (3.189)

$$\begin{aligned}0 &= \sum_{i=1}^{n}\mathbf{K}_i(\phi - t_i) + \mathbf{K}_0(\phi - t_0)\\ &= \mathbf{K}_D(\phi - t_D) + \mathbf{K}_0(\phi - t_0).\end{aligned} \tag{3.207}$$

For existing $\mathbf{K}_0^{-1}$,

$$\phi = t_0 + \mathbf{K}_0^{-1}\mathbf{K}_D(t_D - \phi). \tag{3.208}$$

Hence one can introduce

$$a = \mathbf{K}_D(t_D - \phi) \tag{3.209}$$

to obtain

$$\phi = t_0 + \mathbf{K}_0^{-1} a. \tag{3.210}$$

Eq. (3.210) already shows that $\phi - t_0$ is given by a superposition of the kernel functions $\mathbf{K}_0^{-1}(x, x_i)$ of the prior covariance with one component for each measured x–value. The coefficients a, however, depend on the data and still have to be calculated. Inserting Eq. (3.210) into Eq. (3.209) one finds an equation for a

$$\left(\mathbf{I} + \mathbf{K}_D \mathbf{K}_0^{-1}\right) a = \mathbf{K}_D (t_D - t_0). \tag{3.211}$$

Multiplying Eq. (3.211) from the left by the projector $\mathbf{I}_D$ defined in (3.182) and using

$$\mathbf{K}_D = \mathbf{K}_D \mathbf{I}_D = \mathbf{I}_D \mathbf{K}_D, \quad a = \mathbf{I}_D a, \quad t_D = \mathbf{I}_D t_D, \tag{3.212}$$

one obtains an equation in data space

$$\left(\mathbf{I}_D + \mathbf{K}_D \mathbf{K}_{0,DD}^{-1}\right) a = \mathbf{K}_D (t_D - t_{0,D}), \tag{3.213}$$

where

$$\mathbf{K}_{0,DD}^{-1} = (\mathbf{K}_0^{-1})_{DD} = \mathbf{I}_D \mathbf{K}_0^{-1} \mathbf{I}_D \neq (\mathbf{K}_{0,DD})^{-1}, \quad t_{0,D} = \mathbf{I}_D t_0. \tag{3.214}$$

Thus

$$a = \mathbf{C}_{DD}\, b, \tag{3.215}$$

where

$$\mathbf{C}_{DD} = \left(\mathbf{I}_D + \mathbf{K}_D \mathbf{K}_{0,DD}^{-1}\right)^{-1} \tag{3.216}$$

and

$$b = \mathbf{K}_D (t_D - t_{0,D}). \tag{3.217}$$

In components Eq. (3.213) reads

$$\sum_l \left(\delta_{kl} + n_{x_k} \mathbf{K}_0^{-1}(x_k, x_l)\right) a(x_l) = n_{x_k} \left(t_D(x_k) - t_0(x_k)\right), \tag{3.218}$$

for $\sigma = 1$ and n_{x_k} denoting the number of data points at x_k. Having calculated a, the solution ϕ in terms of the prior covariance $\mathbf{K}_0^{-1}$ is found by inserting in Eq. (3.210)

$$\begin{aligned} \phi &= t_0 + \mathbf{K}_0^{-1} \mathbf{C}_{DD} b \\ &= t_0 + \mathbf{K}_0^{-1} \left(\mathbf{K}_D^{-1} + \mathbf{K}_{0,DD}^{-1}\right)^{-1} (t_D - t_0) \\ &= t_0 + \mathbf{K}_0^{-1} \widetilde{\mathbf{K}} (t_D - t_0), \end{aligned} \tag{3.219}$$

where t_0 can be replaced by $t_{0,D}$ and we defined the $(\tilde{n}\times\tilde{n})$–matrix ($\tilde{n}$ denoting the number of different x_i–values in the training data)

$$\widetilde{\mathbf{K}} = \left(\mathbf{K}_D^{-1} + \mathbf{K}_{0,DD}^{-1}\right)^{-1} = \mathbf{C}_{DD}\mathbf{K}_D. \tag{3.220}$$

Eq. (3.219) can also be obtained directly from Eqs. (3.197) and (3.193), without introducing the auxiliary variable a, using $\mathbf{K} = \mathbf{K}_0 + \mathbf{K}_D$,

$$\begin{aligned} \phi = t &= \mathbf{K}^{-1}\left(\mathbf{K}_0 t_0 + \mathbf{K}_D t_D\right) \\ &= \mathbf{K}^{-1}\left(\mathbf{K}\, t_0 - \mathbf{K}_D t_0 + \mathbf{K}_D t_D\right) \\ &= t_0 + \mathbf{K}^{-1}\mathbf{K}_D\left(t_D - t_0\right) \end{aligned} \tag{3.221}$$

with

$$\begin{aligned} \mathbf{K}^{-1}\mathbf{K}_D &= \mathbf{K}_0^{-1}\left(\mathbf{I} + \mathbf{K}_D\mathbf{K}_0^{-1}\right)^{-1}\mathbf{K}_D \\ &= \mathbf{K}_0^{-1}\left(\mathbf{I} + \mathbf{K}_D\mathbf{K}_{0,DD}^{-1}\right)^{-1}\mathbf{K}_D \\ &= \mathbf{K}_0^{-1}\mathbf{C}_{DD}\mathbf{K}_D \\ &= \mathbf{K}_0^{-1}\widetilde{\mathbf{K}}. \end{aligned} \tag{3.222}$$

Like (3.204) the operator

$$\mathbf{K}_0^{-1}\left(\mathbf{I}_D + \mathbf{K}_D\mathbf{K}_{0,DD}^{-1}\right)^{-1} = \mathbf{K}_0^{-1}\mathbf{C}_{DD}, \tag{3.223}$$

or its extension $\mathbf{K}^{-1}$, is also known as the equivalent kernel, due to its relation to kernel smoothing techniques [162, 198, 203, 455] and its restriction to data space $\mathbf{K}_{DD}^{-1}$ as influence matrix [507]. Indeed, for $t_0 = 0$ the solution $\phi = \mathbf{K}^{-1}\mathbf{K}_D t_D = \mathbf{K}^{-1}\sum_i \mathbf{K}_i t_i$ is just a superposition of kernel functions

$$\phi(x) = \sum_i \mathbf{K}^{-1}(x, x_i)\, y_i, \tag{3.224}$$

with components located at each data point x_i.

Interestingly, Eq. (3.210) still holds for *non–quadratic* data terms of the form $g_D(\phi)$ with any differentiable function fulfilling $g(\phi) = g(\phi_D)$, where $\phi_D = \mathbf{I}_D\phi$ is the restriction of ϕ to data space. Hence also the function of functional derivatives with respect to $\phi(x)$ is restricted to data space, i.e., $g'(\phi_D) = g'_D(\phi_D)$ with $g'_D = \mathbf{I}_D g'$ and $g'(\phi, x) = \delta g(\phi)/\delta\phi(x)$. For example, $g(\phi) = \sum_{i=1}^n V(\phi(x_i) - y_i)$ with V a differentiable function. The finite dimensional vector a is then found by solving a nonlinear equation instead of a linear one [158, 163]. An example of a nondifferentiable V is given by the ϵ–insensitive error (3.281) as used in regression support vector machines (see Section 3.8.4).

Non-invertible $\mathbf{K}_0$: For non-invertible $\mathbf{K}_0$ one can solve for ϕ using the Moore–Penrose inverse $\mathbf{K}_0^{\#}$. Let us first recall some basic facts [32,136,267, 283,352,392,393]. A pseudo inverse $\mathbf{B}$ of (a possibly non-square) $\mathbf{A}$ is defined by the conditions

$$\mathbf{BAB} = \mathbf{B}, \quad \mathbf{ABA} = \mathbf{A}, \tag{3.225}$$

and becomes for real $\mathbf{A}$ the unique Moore–Penrose inverse $\mathbf{A}^{\#}$ if

$$(\mathbf{AA}^{\#})^T = \mathbf{AA}^{\#}, \quad (\mathbf{A}^{\#}\mathbf{A})^T = \mathbf{A}^{\#}\mathbf{A}. \tag{3.226}$$

A linear equation

$$\mathbf{A}x = b \tag{3.227}$$

is solvable if

$$\mathbf{AA}^{\#}b = b. \tag{3.228}$$

In that case the solution is

$$x = \mathbf{A}^{\#}b + x^0 = \mathbf{A}^{\#}b + y - \mathbf{A}^{\#}\mathbf{A}y, \tag{3.229}$$

where $x^0 = y - \mathbf{A}^{\#}\mathbf{A}y$ solves the homogeneous equation $\mathbf{A}x^0 = 0$ and vector y is arbitrary. Hence x^0 can be expanded in an orthonormalized basis ψ_l of the null space of $\mathbf{A}$

$$x^0 = \sum_l c_l \psi_l. \tag{3.230}$$

For a diagonal

$$\mathbf{D} = \begin{pmatrix} \widetilde{\mathbf{D}} & 0 \\ 0 & 0 \end{pmatrix}, \tag{3.231}$$

where $\widetilde{\mathbf{D}}$ is invertible and 0 stands for a zero matrix, the (unique) Moore–Penrose inverse is

$$\mathbf{D}^{\#} = \begin{pmatrix} \widetilde{\mathbf{D}}^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.232}$$

(Note, that it is, for example, not possible to add to $\mathbf{D}^{\#}$ a matrix which is non-zero in the null space of $\mathbf{D}$, as then Eqs. (3.225) are not fulfilled.) For a square matrix $\mathbf{D}$ the product

$$\mathbf{DD}^{\#} = \mathbf{D}^{\#}\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \mathbf{I}_1^D \tag{3.233}$$

becomes the projector on the space where $\mathbf{D}$ is invertible. Similarly, $\mathbf{I}_0^D = \mathbf{I} - \mathbf{I}_1^D$ is the projector on the null space of $\mathbf{D}$.

For a general (possibly non-square) $\mathbf{A}$ there always exist orthogonal $\mathbf{O_1}$ and $\mathbf{O_2}$ so that $\mathbf{A}$ can be written in orthogonal normal form $\mathbf{A} = \mathbf{O_1 D O_2}$, where $\mathbf{D}$ is of the form of Eq. (3.231) [267]. The corresponding Moore–Penrose inverse is $\mathbf{A}^{\#} = \mathbf{O_2^{-1}} \mathbf{D}^{\#} \mathbf{O_1^{-1}}$.

Similarly, for some $\mathbf{A}$ which can be diagonalized, i.e., for a square matrix $\mathbf{A} = \mathbf{M}^{-1}\mathbf{D}\mathbf{M}$ with diagonal $\mathbf{D}$, the Moore–Penrose inverse is $\mathbf{A}^{\#} = \mathbf{M}^{-1}\mathbf{D}^{\#}\mathbf{M}$. From Eq. (3.233) it follows then that

$$\mathbf{A}\mathbf{A}^{\#} = \mathbf{A}^{\#}\mathbf{A} = \mathbf{I}_1 = \mathbf{I} - \mathbf{I}_0, \tag{3.234}$$

where $\mathbf{I}_1 = \mathbf{M}^{-1}\mathbf{D}\mathbf{D}^{\#}\mathbf{M} = \mathbf{M}^{-1}\mathbf{I}_1^D\mathbf{M}$ projects on the space where $\mathbf{A}$ is invertible and $\mathbf{I}_0 = \mathbf{I} - \mathbf{I}_1 = \sum_l \psi_l \psi_l^T$ is the projector into the null space of $\mathbf{A}$. Thus using the definition (3.225) yields

$$\mathbf{I}_1\mathbf{A}\mathbf{I}_1 = \mathbf{A}, \qquad \mathbf{I}_1\mathbf{A}^{\#}\mathbf{I}_1 = \mathbf{A}^{\#}. \tag{3.235}$$

Sometimes it is convenient to write in that case

$$\mathbf{A}^{\#} = (\mathbf{I}_1\mathbf{A}\mathbf{I}_1)^{-1}, \tag{3.236}$$

where the inversion is understood to be performed in the subspace where $\mathbf{A}$ is invertible. Expressed by projectors the solvability condition Eq. (3.228) becomes

$$\mathbf{I}_1 b = b, \quad \text{or} \quad \mathbf{I}_0 b = 0, \tag{3.237}$$

and in terms of a basis ψ_l of the null space

$$\langle \psi_l | \, b \rangle = 0, \ \forall l. \tag{3.238}$$

This means that the inhomogeneity b must have no components within the null space of $\mathbf{A}$. We can then write

$$x^0 = \mathbf{I}_0 \, y, \tag{3.239}$$

showing that a solution x^0 of the homogeneous Equation $\mathbf{A}x^0 = 0$ can be obtained by projecting an arbitrary y on the null space of $\mathbf{A}$. Collecting the results for some $\mathbf{A}$, which can be diagonalized, Eq. (3.229) can be expressed in terms of projection operators as follows, using Eqs. (3.236), (3.237), (3.239),

$$x = \mathbf{A}^{\#}b + \mathbf{I}_0 y = (\mathbf{I}_1\mathbf{A}\mathbf{I}_1)^{-1}\mathbf{I}_1 b + \mathbf{I}_0 \, y. \tag{3.240}$$

Thus a solution x is obtained by inverting $\mathbf{A}$ in the subspace where it is invertible and adding an arbitrary component of the null space.

Now we apply this result to Eq. (3.207), where $\mathbf{K}_0$, being positive semi-definite, is diagonalizable. (In this case $\mathbf{M}$ is an orthogonal matrix and the entries of D are real and larger or equal to zero.) Hence one obtains for the condition Eq. (3.237)

$$\mathbf{I}_0 \left(\mathbf{K}_0 t_0 + \mathbf{K}_D(t_D - \phi)\right) = 0, \tag{3.241}$$

and for Eq. (3.229)

$$\phi = \mathbf{K}_0^{\#} \left(\mathbf{K}_0 t_0 + \mathbf{K}_D(t_D - \phi)\right) + \phi^0, \tag{3.242}$$

where $\mathbf{K}_0\phi^0 = 0$ so that $\phi^0 = \sum_l c_l\psi_l$ can be expanded in an orthonormalized basis ψ_l of the null space of $\mathbf{K}_0$, assumed here to be of finite dimension. To find an equation in data space we use Eq. (3.209)

$$a = \mathbf{K}_D(t_D - \phi), \tag{3.243}$$

to get from Eqs. (3.241) and (3.242)

$$0 = \langle\psi_l|\mathbf{K}_0|t_0\rangle + \langle\psi_l|a\rangle, \quad \forall l \tag{3.244}$$

$$\phi = \mathbf{K}_0^{\#}(\mathbf{K}_0 t_0 + a) + \sum_l c_l\psi_l. \tag{3.245}$$

These equations have to be solved for a and the coefficients c_l. Inserting Eq. (3.245) into the definition (3.243) gives, using $\mathbf{K}_0^{\#}\mathbf{K}_0 = \mathbf{I}_1$ according to Eq. (3.234),

$$(\mathbf{I} + \mathbf{K}_D\mathbf{K}_0^{\#})a = \mathbf{K}_D t_D - \mathbf{K}_D\mathbf{I}_1 t_0 - \mathbf{K}_D\sum_l c_l\psi_l. \tag{3.246}$$

Eq. (3.246) generalizes Eq. (3.211) to the case of non-invertible $\mathbf{K}_0$. With $a = \mathbf{I}_D a$ (where $\mathbf{I}_D$, the projector on the space of training data, is defined in (3.182)) the solvability condition (3.241), or equivalently (3.244), becomes

$$\sum_{i=1}^{\tilde{n}} \psi_l(x_i)\, a(x_i) = -\langle\psi_l|\mathbf{K}_0|t_0\rangle, \quad \forall l, \tag{3.247}$$

the sum going over different x_i only. Eq. (3.246) for a and c_l reads in data space, similar to Eq. (3.213),

$$a = \tilde{\mathbf{C}}\tilde{b}, \tag{3.248}$$

where $\tilde{\mathbf{C}}^{-1} = \mathbf{I} + \mathbf{K}_D\mathbf{K}_0^{\#}$ has been assumed invertible and $\tilde{b}$ is given by the right hand side of Eq. (3.246). Inserting into Eq. (3.245) the solution finally can be written

$$\phi = \mathbf{I}_1 t_0 + \mathbf{K}_0^{\#}\tilde{\mathbf{C}}\tilde{b} + \sum_l c_l\psi_l, \tag{3.249}$$

where c_l has to be determined by combining Eq. (3.249) with Eq. (3.247).

Again, general non-quadratic data terms $g(\phi_D)$ can be allowed. In that case $\delta g(\phi_D)/\delta\phi(x) = g'(\phi_D, x) = (\mathbf{I}_D g')(\phi_D, x)$ and Eq. (3.243) becomes the nonlinear equation

$$a = g'(\phi_D) = g'\Big(\mathbf{I}_D\left(\mathbf{K}_0^{\#}(\mathbf{K}_0 t_0 + \mathbf{K}_D(t_D - \phi)) + \phi^0\right)\Big). \tag{3.250}$$

The solution a of Eq. (3.250) has then to be inserted in Eq. (3.245).

3.8.2 Exact predictive density

For Gaussian regression the predictive density under training data D and prior D_0 can be found analytically without resorting to a saddle point approximation. The predictive density has been defined in Eq. (2.64) as the ϕ-integral

$$\begin{aligned} p(y|x,D,D_0) &= \int d\phi\, p(y|x,\phi)p(\phi|D,D_0) \\ &= \frac{\int d\phi\, p(y|x,\phi)p(y_D|x_D,\phi)p(\phi|D_0)}{\int d\phi\, p(y_D|x_D,\phi)p(\phi|D_0)} \\ &= \frac{p(y,y_D|x,x_D,D_0)}{p(y_D|x_D,D_0)}. \end{aligned} \tag{3.251}$$

Denoting training data values y_i by t_i sampled with inverse covariances $\mathbf{K}_i$ at x_i and analogously test data values $y = y_{n+1}$ by t_{n+1} sampled with inverse covariance $\mathbf{K}_{n+1}$, we have for $1 \leq i \leq n+1$

$$p(y_i|x_i,\phi) = \det(\mathbf{K}_i/2\pi)^{\frac{1}{2}} e^{-\frac{1}{2}\langle \phi - t_i|\mathbf{K}_i|\phi - t_i\rangle} \tag{3.252}$$

and

$$p(\phi|D_0) = \det(\mathbf{K_0}/2\pi)^{\frac{1}{2}} e^{-\frac{1}{2}\langle \phi - t_0|\mathbf{K}_0|\phi - t_0\rangle}, \tag{3.253}$$

hence

$$p(y|x,D,D_0) = \frac{\int d\phi\, e^{-\frac{1}{2}\sum_{i=0}^{n+1}\langle \phi - t_i|\mathbf{K}_i|\phi - t_i\rangle + \frac{1}{2}\sum_{i=0}^{n+1}\ln \det_i(\mathbf{K}_i/2\pi)}}{\int d\phi\, e^{-\frac{1}{2}\sum_{i=0}^{n}\langle \phi - t_i|\mathbf{K}_i|\phi - t_i\rangle + \frac{1}{2}\sum_{i=0}^{n}\ln \det_i(\mathbf{K}_i/2\pi)}}. \tag{3.254}$$

Here we have written $\det_i(\mathbf{K}_i/2\pi)$ with a subscript i for a determinant calculated in the space where $\mathbf{K}_i$ is invertible. This notation is useful because in general $\det_i \mathbf{K}_i \det \mathbf{K}_0 \neq \det_i \mathbf{K}_i\mathbf{K}_0$. Using the generalized 'bias–variance' decomposition (3.160) yields

$$p(y|x,D,D_0) = \frac{\int d\phi\, e^{-\frac{1}{2}\langle \phi - t_+|\mathbf{K}_+|\phi - t_+\rangle - \frac{n+2}{2}V_+ + \frac{1}{2}\sum_{i=0}^{n+1}\ln \det_i(\mathbf{K}_i/2\pi)}}{\int d\phi\, e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t\rangle - \frac{n+1}{2}V + \frac{1}{2}\sum_{i=0}^{n}\ln \det_i(\mathbf{K}_i/2\pi)}}, \tag{3.255}$$

with

$$t = \mathbf{K}^{-1}\sum_{i=0}^{n}\mathbf{K}_i t_i, \qquad \mathbf{K} = \sum_{i=0}^{n}\mathbf{K}_i, \tag{3.256}$$

$$t_+ = \mathbf{K}_+^{-1}\sum_{i=0}^{n+1}\mathbf{K}_i t_i, \qquad \mathbf{K}_+ = \sum_{i=0}^{n+1}\mathbf{K}_i, \tag{3.257}$$

$$V = \frac{1}{n+1}\sum_{i=0}^{n}\langle t_i|\mathbf{K}_i|t_i\rangle - \left\langle t\left|\frac{\mathbf{K}}{n+1}\right|t\right\rangle, \tag{3.258}$$

$$V_+ = \frac{1}{n+2}\sum_{i=0}^{n+1}\langle t_i|\mathbf{K}_i|t_i\rangle - \left\langle t_+\left|\frac{\mathbf{K}_+}{n+2}\right|t_+\right\rangle. \tag{3.259}$$

Now the ϕ–integration can be performed

$$p(y|x,D,D_0) = \frac{e^{-\frac{n+2}{2}V_+ + \frac{1}{2}\sum_{i=0}^{n+1}\ln\det_i(\mathbf{K}_i/2\pi) - \frac{1}{2}\ln\det(\mathbf{K}_+/2\pi)}}{e^{-\frac{n+1}{2}V + \frac{1}{2}\sum_{i=0}^{n}\ln\det_i(\mathbf{K}_i/2\pi) - \frac{1}{2}\ln\det(\mathbf{K}/2\pi)}}. \tag{3.260}$$

Canceling common factors, writing again y for t_{n+1}, $\mathbf{K}_x$ for $\mathbf{K}_{n+1}$, $\det_x$ for $\det_{n+1}$, and using $\mathbf{K}_+ t_+ = \mathbf{K}t + \mathbf{K}_x y$, Eq. (3.260) becomes

$$p(y|x,D,D_0) = e^{-\frac{1}{2}\langle y|\mathbf{K}_y|y\rangle + \langle y|\mathbf{K}_y|t\rangle + \frac{1}{2}\langle t|\mathbf{K}\mathbf{K}_+^{-1}\mathbf{K}-\mathbf{K}|t\rangle + \frac{1}{2}\ln\det_x(\mathbf{K}_x\mathbf{K}_+^{-1}\mathbf{K}/2\pi)}. \tag{3.261}$$

Here we introduced $\mathbf{K}_y = \mathbf{K}_y^T = \mathbf{K}_x - \mathbf{K}_x\mathbf{K}_+^{-1}\mathbf{K}_x$ and used that

$$\det\mathbf{K}^{-1}\mathbf{K}_+ = \det(\mathbf{I} + \mathbf{K}^{-1}\mathbf{K}_x) = \det{}_x\mathbf{K}^{-1}\mathbf{K}_+ \tag{3.262}$$

can be calculated in the space of test data x. This follows from $\mathbf{K} = \mathbf{K}_+ - \mathbf{K}_x$ and the equality

$$\det\begin{pmatrix} 1+A_{xx} & 0 \\ B_{\bar{x}x} & 1 \end{pmatrix} = \det(1+A_{xx}), \tag{3.263}$$

with $A_{xx} = \mathbf{I}_x\mathbf{K}^{-1}\mathbf{K}_x$, $B_{\bar{x}x} = (\mathbf{I}-\mathbf{I}_x)\mathbf{K}^{-1}\mathbf{K}_x$, and $\mathbf{I}_x$ denoting the projector into the space of test data x. Finally,

$$\mathbf{K}_y = \mathbf{K}_x - \mathbf{K}_x\mathbf{K}_+^{-1}\mathbf{K}_x = \mathbf{K}_x\mathbf{K}_+^{-1}\mathbf{K} = (\mathbf{K} - \mathbf{K}\mathbf{K}_+^{-1}\mathbf{K}) \tag{3.264}$$

yields the correct normalization of the predictive density

$$p(y|x,D,D_0) = e^{-\frac{1}{2}\langle y-\bar{y}|\mathbf{K}_y|y-\bar{y}\rangle + \frac{1}{2}\ln\det_x(\mathbf{K}_y/2\pi)}, \tag{3.265}$$

with mean and (co–)variance

$$\bar{y} = t = \mathbf{K}^{-1}\sum_{i=0}^{n}\mathbf{K}_i t_i, \tag{3.266}$$

$$\mathbf{K}_y^{-1} = \left(\mathbf{K}_x - \mathbf{K}_x\mathbf{K}_+^{-1}\mathbf{K}_x\right)^{-1} = \mathbf{K}_x^{-1} + \mathbf{I}_x\mathbf{K}^{-1}\mathbf{I}_x. \tag{3.267}$$

It is useful to express the posterior covariance $\mathbf{K}^{-1}$ by the prior covariance $\mathbf{K}_0^{-1}$. According to

$$\begin{pmatrix} 1+A_{DD} & B_{D\bar{D}} \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} (1+A_{DD})^{-1} & -(1+A_{DD})^{-1}B_{D\bar{D}} \\ 0 & 1 \end{pmatrix}, \tag{3.268}$$

with $A_{DD} = \mathbf{K}_D\mathbf{K}_{0,DD}^{-1}$, $B_{D\bar{D}} = \mathbf{K}_D\mathbf{K}_{0,D\bar{D}}^{-1}$, where $\mathbf{K}_{0,DD}^{-1} = \mathbf{I}_D\mathbf{K}_0^{-1}\mathbf{I}_D$, $\mathbf{K}_{0,D\bar{D}}^{-1} = \mathbf{I}_D\mathbf{K}_0^{-1}\mathbf{I}_{\bar{D}}$, and $\mathbf{I}_{\bar{D}} = \mathbf{I} - \mathbf{I}_D$, $\mathbf{I}_D$ denoting the projector into the space of training data, we find, analogous to Eq. (3.222),

$$\begin{aligned}\mathbf{K}^{-1} &= \mathbf{K}_0^{-1}\left(\mathbf{I} + \mathbf{K}_D\mathbf{K}_0^{-1}\right)^{-1} \\ &= \mathbf{K}_0^{-1}\Bigg(\left(\mathbf{I}_D + \mathbf{K}_D\mathbf{K}_{0,DD}^{-1}\right)^{-1} \\ &\quad - \left(\mathbf{I}_D + \mathbf{K}_D\mathbf{K}_{0,DD}^{-1}\right)^{-1}\mathbf{K}_D\mathbf{K}_{0,D\bar{D}}^{-1} + \mathbf{I}_{\bar{D}}\Bigg).\end{aligned} \tag{3.269}$$

Notice that while $\mathbf{K}_D^{-1} = (\mathbf{I}_D\mathbf{K}_D\mathbf{I}_D)^{-1}$, we have in general $\mathbf{K}_{0,DD}^{-1} = \mathbf{I}_D\mathbf{K}_0^{-1}\mathbf{I}_D \neq (\mathbf{I}_D\mathbf{K}_0\mathbf{I}_D)^{-1}$. This means that $\mathbf{K}_0^{-1}$ has to be known to find $\mathbf{K}_{0,DD}^{-1}$ and it is not enough to invert $\mathbf{I}_D\mathbf{K}_0\mathbf{I}_D = \mathbf{K}_{0,DD} \neq (\mathbf{K}_{0,DD}^{-1})^{-1}$. In data space

$$\left(\mathbf{I}_D + \mathbf{K}_D\mathbf{K}_{0,DD}^{-1}\right)^{-1} = \left(\mathbf{K}_D^{-1} + \mathbf{K}_{0,DD}^{-1}\right)^{-1}\mathbf{K}_D^{-1}, \tag{3.270}$$

hence Eq. (3.269) results in the space of non–training data in

$$\mathbf{K}_{\bar{D}\bar{D}}^{-1} = \mathbf{K}_{0,\bar{D}\bar{D}}^{-1} - \mathbf{K}_{0,\bar{D}D}^{-1}\left(\mathbf{K}_D^{-1} + \mathbf{K}_{0,DD}^{-1}\right)^{-1}\mathbf{K}_{0,D\bar{D}}^{-1}. \tag{3.271}$$

This allows now to express for $x \notin x_D$ the predictive mean (3.266) and the (co–)variance (3.267) by the prior covariance

$$\bar{y} = t_0 + \mathbf{K}_0^{-1}\left(\mathbf{K}_D^{-1} + \mathbf{K}_{0,DD}^{-1}\right)^{-1}(t_D - t_0), \tag{3.272}$$

$$\mathbf{K}_y^{-1} = \mathbf{K}_x^{-1} + \mathbf{K}_{0,xx}^{-1} - \mathbf{K}_{0,xD}^{-1}\left(\mathbf{K}_D^{-1} + \mathbf{K}_{0,DD}^{-1}\right)^{-1}\mathbf{K}_{0,Dx}^{-1}, \tag{3.273}$$

where $\mathbf{K}_{0,Dx}^{-1} = \mathbf{I}_D\mathbf{K}_0^{-1}\mathbf{I}_\mathbf{x}$ and Eq. (3.272) for $\bar{y}$ corresponds to Eq. (3.219). For given prior covariance $\mathbf{K}_0^{-1}$, both $\bar{y}$ and $\mathbf{K}_y^{-1}$ can be obtained by calculating the $(\tilde{n} \times \tilde{n})$–matrix of Eq. (3.220)

$$\widetilde{\mathbf{K}} = \left(\mathbf{K}_D^{-1} + \mathbf{K}_{0,DD}^{-1}\right)^{-1}. \tag{3.274}$$

Comparison of Eqs. (3.272) and (3.273) with the maximum posterior solution ϕ^* of Eq. (3.219) now shows that for Gaussian regression the exact predictive density $p(y|x, D, D_0)$ and its maximum posterior approximation $p(y|x, \phi^*)$ have the same mean

$$t = \int dy\, y\, p(y|x, D, D_0) = \int dy\, y\, p(y|x, \phi^*). \tag{3.275}$$

The variances, however, differ according to Eq. (3.267) by the term $\mathbf{I}_x\mathbf{K}^{-1}\mathbf{I}_x$.

According to the results of Section 2.2.2 the mean of the predictive density is the optimal choice under squared–error loss (2.158). For Gaussian regression, the optimal regression function $a^*(x)$ is therefore the same for squared–error loss in exact and in maximum posterior treatment and thus according to Section 2.2.2 also for log–loss (being equal to squared–error loss for Gaussian $p(y|x,a)$ with fixed variance),

$$a^*_{\text{MAP,log}} = a^*_{\text{exact,log}} = a^*_{\text{MAP,sq.}} = a^*_{\text{exact,sq.}} = \phi^* = t. \tag{3.276}$$

In case the space of possible $p(y|x,a)$ is not restricted to Gaussian densities with fixed variance, the variance of the optimal density under log–loss $p(y|x,a^*_{\text{exact,log}}) = p(y|x,D,D_0)$ differs by $\mathbf{I}_x\mathbf{K}^{-1}\mathbf{I}_x$ from its maximum posterior approximation $p(y|x,a^*_{\text{MAP,log}}) = p(y|x,\phi^*)$.

3.8.3 Gaussian mixture regression (cluster regression)

A generalized version of Gaussian regression is obtained by allowing the likelihoods to become mixtures of Gaussians

$$p(y|x,\phi) = \frac{\sum_k^m p(k)\, e^{-\frac{\beta}{2}|y-\phi_k(x)|^2}}{\int dy \sum_k^m p(k)\, e^{-\frac{\beta}{2}|y-\phi_k(x)|^2}}, \tag{3.277}$$

where m denotes the number of mixture components ('cluster') and the normalization factor is $\sum_k p(k)\left(\frac{\beta}{2\pi}\right)^{\frac{m}{2}}$. Hence ϕ is specified by mixing coefficients $p(k)$ and a vector of regression functions $\phi_k(x)$ specifying the x–dependent location of the kth cluster centroid of the mixture model. A simple prior for $\phi_k(x)$ is a smoothness prior diagonal in the cluster components. As any density $p(y|x,\phi)$ can be approximated arbitrarily well by a mixture with large enough m, such cluster regression models allow to interpolate between Gaussian regression and general density estimation.

The posterior density becomes for independent data

$$p(\phi|D,D_0) = \frac{p(\phi|D_0)}{p(y_D|x_D,D_0)} \prod_i^n \frac{\sum_k^m p(k)\, e^{-\frac{\beta}{2}(y_i-\phi_k(x_i))^2}}{\sum_k^m p(k)\left(\frac{\beta}{2\pi}\right)^{\frac{m}{2}}}. \tag{3.278}$$

Maximizing that posterior, for fixed x, uniform $p(k)$ and $p(\phi|D_0)$, is equivalent to the *clustering* approach of Rose, Gurewitz, and Fox for squared distance costs [428].

In general, mixture models for likelihoods can be used to extend all kinds of restricted likelihood models [123, 271, 324, 485]. They have been used also in a variety of different contexts besides regression, for instance, to extend graphical models [246, 335] or in the analysis of experimental spectra [130].

3.8.4 Support vector machines and regression

Expanding the regression function $\phi(x)$ in a basis of eigenfunctions Ψ_k of $\mathbf{K}_0$

$$K_0 = \sum_k \lambda_k \Psi_k \Psi_k^T, \quad \phi(x) = \sum_k n_k \Psi_k(x) \tag{3.279}$$

yields for functional (3.186) (and $t_0 \equiv 0$)

$$E(\phi) = \frac{1}{2} \sum_i \left(\sum_k n_k \Psi_k(x_i) - y_i \right)^2 + \sum_k \lambda_k |n_k|^2. \tag{3.280}$$

Under the assumption of output noise for training data, the data terms may, for example, be replaced by the logarithm of a mixture of Gaussians. Such mixture functions with varying mean can develop flat regions where the error is insensitive (robust) to changes of ϕ. Similar to Gaussian mixtures, the mean-squared error data term $|y_i - \phi(x_i)|^2$ may be replaced by an ϵ–insensitive error (or loss) function

$$\begin{aligned} E_i(\phi) &= |y_i - \phi(x_i)|_\epsilon \\ &= \Theta(|y_i - \phi(x_i)| - \epsilon)(|y_i - \phi(x_i)| - \epsilon) \\ &= \begin{cases} 0, & \text{if } |y_i - \phi(x_i)| \le \epsilon, \\ |y_i - \phi(x_i)| - \epsilon, & \text{if } |y_i - \phi(x_i)| > \epsilon, \end{cases} \end{aligned} \tag{3.281}$$

with step function $\Theta(x)$ and a constant ϵ, which is zero for absolute errors smaller ϵ and linear for larger absolute errors (see Fig. 3.10). As the ϵ–insensitive error $|y_i - \phi(x_i)|_\epsilon$ grows in the tails only linearly, it is for large errors again flatter ('more robust') than a mean squared error.[5] The ϵ–insensitive error (3.281) has been expressed as a superposition of Gaussian processes with different variances and means in [405]. Similar error (or loss) functions are the quadratic ϵ–insensitive error [494]

$$E_i(\phi) = |y_i - \phi(x_i)|_\epsilon^2, \tag{3.282}$$

and Huber's loss function [224, 494]

$$E_i(\phi) = \begin{cases} \frac{1}{2}|y_i - \phi(x_i)|^2, & \text{if } |y_i - \phi(x_i)| \le \epsilon, \\ \epsilon|y_i - \phi(x_i)| - \frac{\epsilon^2}{2}, & \text{if } |y_i - \phi(x_i)| > \epsilon. \end{cases} \tag{3.283}$$

Technically, choosing an ϵ–insensitive error (3.281) is convenient because it results in a quadratic programming problem. It is equivalent to Vapnik's support vector machine as derived from its structural risk minimization approach [69, 85, 93, 109, 124, 160, 208, 449, 450, 461–464, 493, 494]. For a discussion of the relation between support vector machines and Gaussian processes see also [378, 441, 442, 466–468, 508, 520].

[5] For robust statistics see for example [55, 177, 196, 213, 224–226, 229, 432, 442, 469, 494, 497].

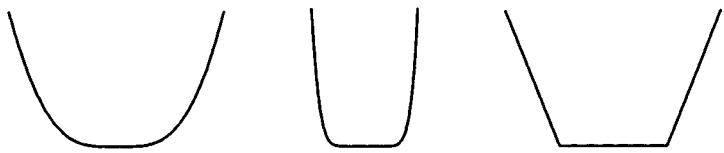


Figure 3.10: Three robust error functions which are insensitive to small errors. Left: Logarithm of mixture with 2 Gaussians with equal variance and different means. Middle: Logarithm of mixture with 11 Gaussians with equal variance and different means. Right: ϵ–insensitive error (3.281). The ϵ–insensitive error also grows only linearly in the tails. Error functions having flat tails (e.g., compared to a quadratic function) are robust with respect to large deviations (outliers) (see also Figs. 5.2a and 5.2b).

3.8.5 Numerical case study: Approximately invariant regression (AIR)

Having now all necessary formulae at hand we are able to study the reconstruction of approximately invariant functions numerically [300]. The Bayesian approach to Gaussian regression, implemented in the following, is schematized in Fig. 3.11: The unknown but fixed true regression function $\phi_{\text{true}}(x)$ is assumed to be sampled from a Gaussian prior process. From the corresponding Gaussian likelihood $p(y|x, \phi_{\text{true}})$ of the form of Eq. (3.164) or (2.11) with σ fixed, training data $D = \{(x_i, y_i)|1 \leq i \leq n\} = \{(x_D, y_D)\}$ are sampled. Given training data and prior, the Bayesian reconstruction of the regression function $\phi(x)$ is then found from Eqs. (3.197) and (3.193) if (the data–dependent) $\mathbf{K}$ can be inverted. Equivalently, if (the data–independent) prior covariance $\mathbf{K}_0^{-1}$ exists and is known, the solution can be obtained from Eq. (3.219). If $\mathbf{K}_0$ is not invertible but its Moore–Penrose inverse $\mathbf{K}_0^{\#}$ can be calculated, Eqs. (3.249) and (3.247) can be used. For the following numerical demonstrations of AIR, true regression functions $\phi(x)$ (discretized on a mesh with 80 points) have been sampled from Gaussian priors of the form (2.106) with zero prior mean $t_0(x)$ and inverse prior covariances $\mathbf{K}_0^{-1}$ implementing either smoothness (approximate infinitesimal translational invariance), approximate periodicity, or an approximately fractal structure.

As reference examples we first study two standard smoothness priors with a negative Laplacian inverse prior covariance

$$\mathbf{K}_0 = -\lambda\Delta \tag{3.284}$$

and a (truncated) Radial Basis Function (RBF) prior [124,291,399,400,507]

$$\mathbf{K}_0 = \lambda \sum_{k=0}^{3} \frac{1}{k!} \left(-\frac{\sigma_0^2 \Delta}{2}\right)^k , \tag{3.285}$$

with hyperparameters λ and σ_0, approximating the 'full' RBF prior $\mathbf{K}_0 = \lambda e^{-\sigma_0^2 \Delta/2}$ of Eq. (2.109). Figs. 3.12 and 3.13 show one typical realization of a sampled regression function $\phi(x)$ for the two prior processes (3.285) and (3.284) with fixed boundary conditions $\phi(1) = \phi(80) = 0$. As is well known, the truncated RBF prior produces smoother regression functions than a negative Laplacian prior. (For fixed boundary values on both, the continuum limit corresponds to a conditional Wiener process whose realizations are continuous but with probability one almost everywhere not differentiable. For a (truncated) RBF prior also higher order derivatives exist, resulting in 'smoother' realizations of ϕ.)

Training data (represented in the figures by the black dots) have been sampled from the random realizations of $\phi(x)$ shown in the figures. Assuming a zero template $t_0 \equiv 0$ and invertible $\mathbf{K}_0$, the reconstructed $\phi(x)$ can be written, according to Eq. (3.193) or Eq. (3.224), as a superposition of the data–dependent equivalent kernel functions $\mathbf{K}^{-1}(x, x_i)$. (In the general case of non–zero t_0 and non–invertible $\mathbf{K}_0$, Eq. (3.249) replaces Eq. (3.224).) Equivalently, as the two inverse prior covariances $\mathbf{K}_0$ are invertible for the given boundary conditions, the reconstructed regression function can be written as a superposition of prior kernel functions $\mathbf{K}_0^{-1}(x, x_i)$ located at the data locations x_i with data–dependent weighting factors as given in Eq. (3.219). The corresponding prior kernel functions $\mathbf{K}_0^{-1}(x, x_i)$ and equivalent kernel functions $\mathbf{K}^{-1}(x, x_i)$ are depicted in Figs. 3.14 and 3.15. For a negative Laplacian prior the reconstructed regression functions are piecewise linear, while for a 'full' RBF prior the reconstruction is a superposition of Gaussians. Reconstructions with a negative Laplacian prior for varying λ are presented in Fig. 3.16. A measure for the quality of the reconstruction is the generalization error being the expectation of the mean squared error on new data

$$E_G = \int dx\, dy\, p(x) p(y|x, \phi_{\text{true}})\, |y - \phi(x)|^2, \tag{3.286}$$

where $p(x)$ denotes the prior density for x. In our case where the true model is known the exact generalization error E_G can be calculated. In practice, E_G has to be estimated using, for example, cross–validation techniques. (Bayesian methods to determine λ are discussed in Section 5.4.5.) Fig. 3.16 shows that in this case using the original λ =1 for reconstruction is optimal. Choosing a larger λ increases the influence of the prior and diminishes the influence of the training data yielding solutions which are too smooth. Going to very small λ leads to the well known overfitting phenomenon where the training data are reproduced nearly exactly (so the training error approaches zero) but the generalization error, measuring the quality of predictions, increases. The reconstruction of $\phi(x)$ for a truncated RBF prior (3.285) is presented in Fig. 3.17 for varying λ and, analogously, in Fig. 3.18, for varying σ_0. For small λ, overfitting is signaled by 'overshooting' in regions without training data, while for σ_0 becoming too small the reconstruction approaches a superposition

of δ–peaks.

As discussed in Section 3.6 the reconstruction of general, approximately invariant functions is technically analogous to the reconstruction of smooth functions, the only difference being the choice of a different prior covariance. As a first example of AIR let us study the reconstruction of functions which are smooth and approximately periodic, i.e., approximately invariant under infinitesimal and specific finite translations. We have seen in Section 3.6.3 that such functions can be generated by a prior process with inverse covariance of the form (3.145), if we choose to implement smoothness by a negative Laplacian prior. A typical regression function sampled from such a prior process is shown in Fig. 3.19. Similarly, implementing smoothness by a truncated RBF prior of the form (3.285) results in regression functions like the one shown in Fig. 3.20. The corresponding prior kernel and equivalent kernel functions clearly show the periodic structure (see Figs. 3.21 and 3.22). Due to the influence of the data the equivalent kernel tends to be more localized than the prior kernel. This effect is clearly seen in Fig. 3.22. Reconstructions with varying regularization parameter can be found in Fig. 3.23 and Fig. 3.24.

Finally, a typical approximately fractal regression function sampled with inverse prior covariance $\mathbf{K}_0 = (1-\gamma)\mathbf{K}_{F3} - \gamma\Delta$ with $\gamma = 0.1$ and $\mathbf{K}_{F3}$ given in (3.154) in Section 3.6.4 can be seen in Fig. 3.25. The approximate fractal structure of the corresponding prior kernel and equivalent kernel function is evident from Fig. 3.26. An example of a reconstruction of an approximate fractal is given in Fig. 3.27.

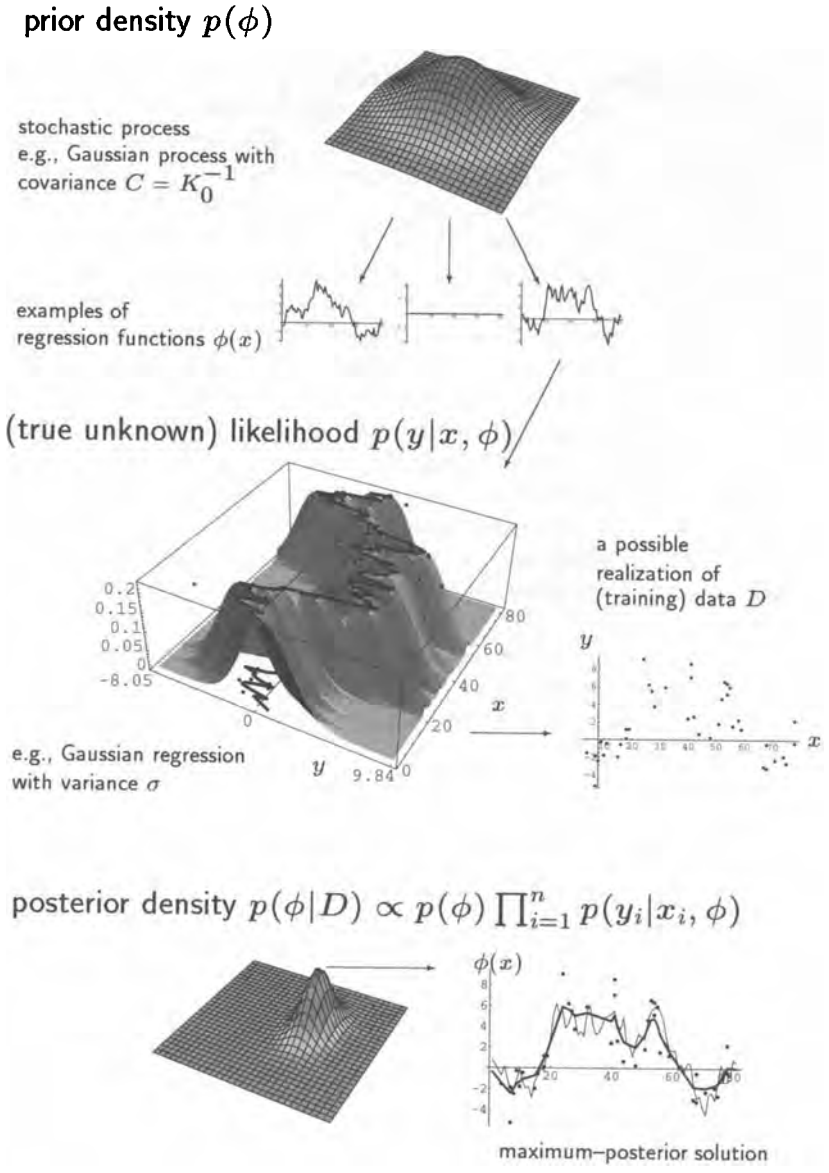


Figure 3.11: AIR 1: Scheme of nonparametric (approximately invariant) Bayesian regression. The 2–dimensional figures of the prior and posterior densities are symbolic representations of stochastic processes over regression functions. The true but unknown likelihood $p(y|x,\phi)$ is assumed Gaussian in y. For comparison, the 3D–picture of the likelihood also shows the corresponding regression function $\phi(x)$ (thick line) and some possible data points (x_i, y_i) (dots) in the top and the bottom plane. The last figure shows the maximum of the posterior, representing the reconstructed regression function (thick line), the corresponding data points (dots), and the original true regression function (thin line).

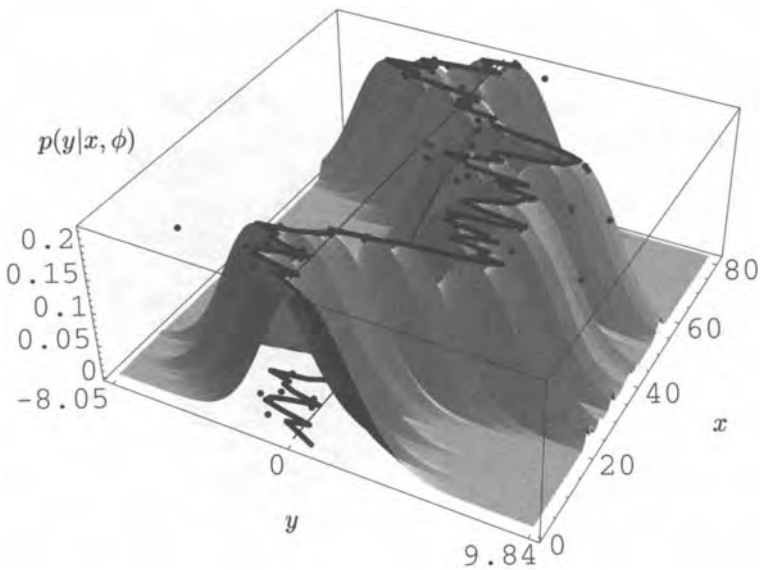


Figure 3.12: AIR 2: Smooth Gaussian likelihood (negative Laplacian). Likelihood $p(y|x,\phi)$ of the form (2.11) with $\sigma = 1.8$ and regression function $\phi(x)$ (thick line) sampled from a Gaussian prior process (2.180) with $t_0(x) \equiv 0$ and negative Laplacian inverse covariance $\mathbf{K}_0 = -\Delta$ implementing approximate invariance under infinitesimal translations (smoothness). (For fixed boundary conditions this corresponds in the continuum limit to a conditional Wiener process, i.e., a Brownian path with fixed endpoints. Its realizations, the sample functions $\phi(x)$, are with probability one almost everywhere not differentiable [507]. The function $\phi(x)$ reconstructed from data D is, however, piecewise linear.) The black dots represent the data used for the reconstruction of $\phi(x)$.

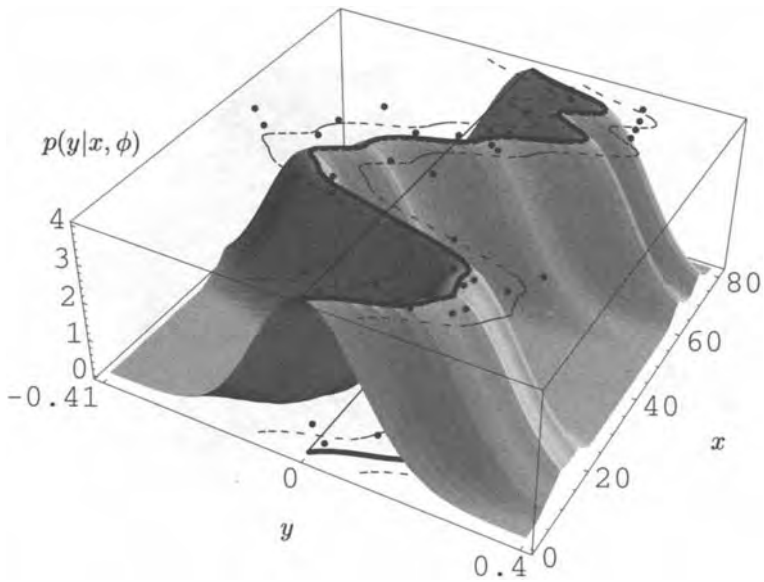


Figure 3.13: AIR 3: Smooth Gaussian likelihood (truncated RBF). Likelihood $p(y|x,\phi)$ of the form (2.11) with $\sigma = 0.1$ and regression function $\phi(x)$ (thick line) sampled from a Gaussian prior process (2.180) with $t_0(x) \equiv 0$ and truncated Radial Basis Function (RBF) inverse covariance $\mathbf{K}_0 = \sum_{k=0}^{3} \frac{1}{k!} \left(-\sigma_0^2 \Delta/2\right)^k \approx e^{-\sigma_0^2 \Delta/2}$, with $\sigma_0 = 10.0$. Compared to the negative Laplacian inverse covariance of Fig. 3.12 typical regression functions $\phi(x)$ look smoother. (Using the full RBF inverse covariance $e^{-\sigma_0^2 \Delta/2}$ all derivatives of $\phi(x)$ exist in the continuum limit.) Besides some data points (dots), the regression function $\phi(x)$ (thick line) the range of one standard deviation above and below the regression function, i.e., $\phi(x) \pm \sigma$, is shown by the two dashed thin lines.

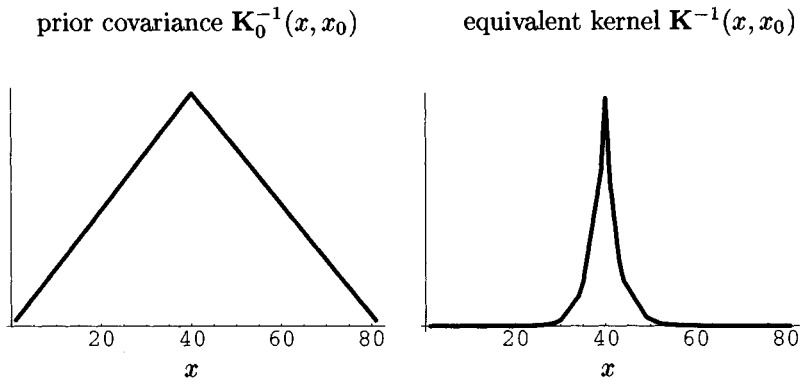


Figure 3.14: AIR 4: Smoothness (negative Laplacian): Prior covariance and equivalent kernel. On the left: Matrix elements $\mathbf{K}_0^{-1}(x, x_0)$ (prior kernel function) for $x_0 = 40$ for the negative Laplacian inverse prior covariance (3.284) with $\lambda = 1$. On the right: The corresponding data–dependent matrix elements $\mathbf{K}^{-1}(x, x_0)$ (equivalent kernel functions) for $x_0 = 40$ and the training data shown in Fig. 3.12.

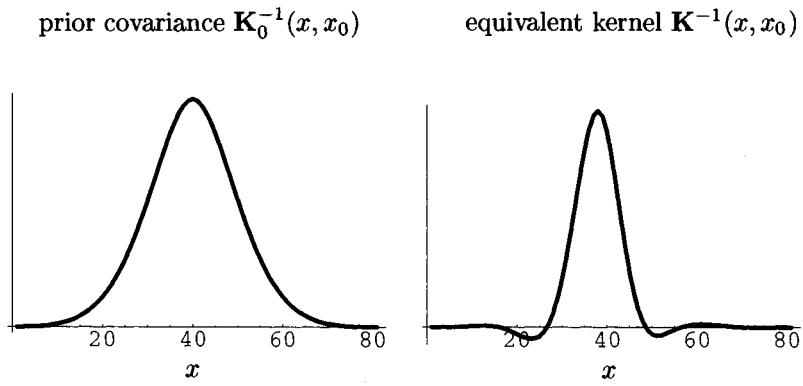


Figure 3.15: AIR 5: Smoothness (truncated RBF): Prior covariance and equivalent kernel. Like Fig. 3.14 but for a truncated RBF prior as in (3.285) and the training data shown in Fig. 3.13.

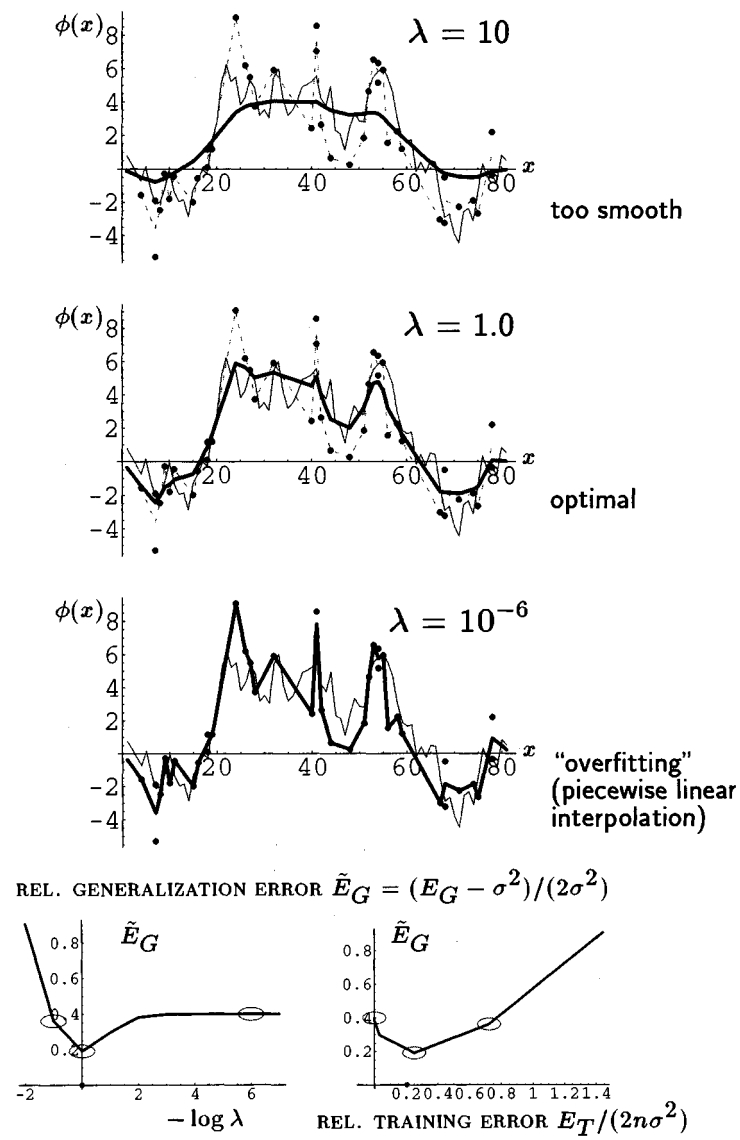


Figure 3.16: AIR 6: Regression with a negative Laplacian smoothness prior (3.284) for the regression (mean) function $\phi(x)$ defining a Gaussian likelihood $p(y|x,\phi)$ with $\sigma = 1.8$. The two figures at the bottom show the dependency of the relative generalization error from $-\log\lambda$ and from the training error. The ovals indicate the positions of the three depicted examples. The training error is defined as $E_T = \sum_{i=1}^{n} |y_i - \phi(x_i)|^2$, the relative training error as $E_T/(2n\sigma^2)$, where n denotes the number of training data, the generalization error as $E_G = \int dx\, dy\, p(x)p(y|x,\phi_{\text{true}})\, |y - \phi(x)|^2$, with uniform $p(x)$, and the relative generalization error as $\tilde{E}_G = (E_G - \sigma^2)/(2\sigma^2)$.

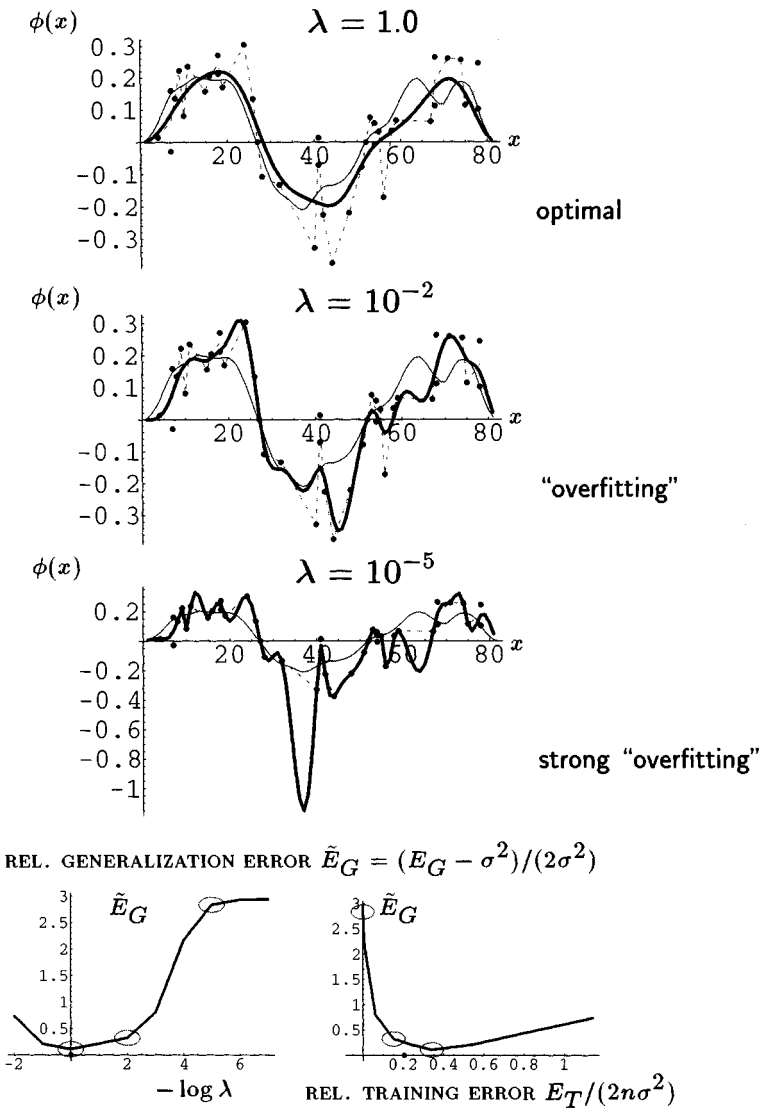


Figure 3.17: AIR 7: Regression with a truncated Radial Basis Function smoothness prior (3.285) $K_0 = \lambda \sum_{k=0}^{3} \frac{1}{k!} \left(-\sigma_0^2 \Delta/2\right)^k \approx \lambda e^{-\sigma_0^2 \Delta/2}$ with σ_0 = 10 and varying regularization factor λ. A Gaussian likelihood $p(y|x,\phi)$ has been chosen with standard deviation $\sigma = 1.8$ and a regression function $\phi_{\text{true}}(x)$ sampled according to the above prior. Training error and generalization error defined as in Fig. 3.16.

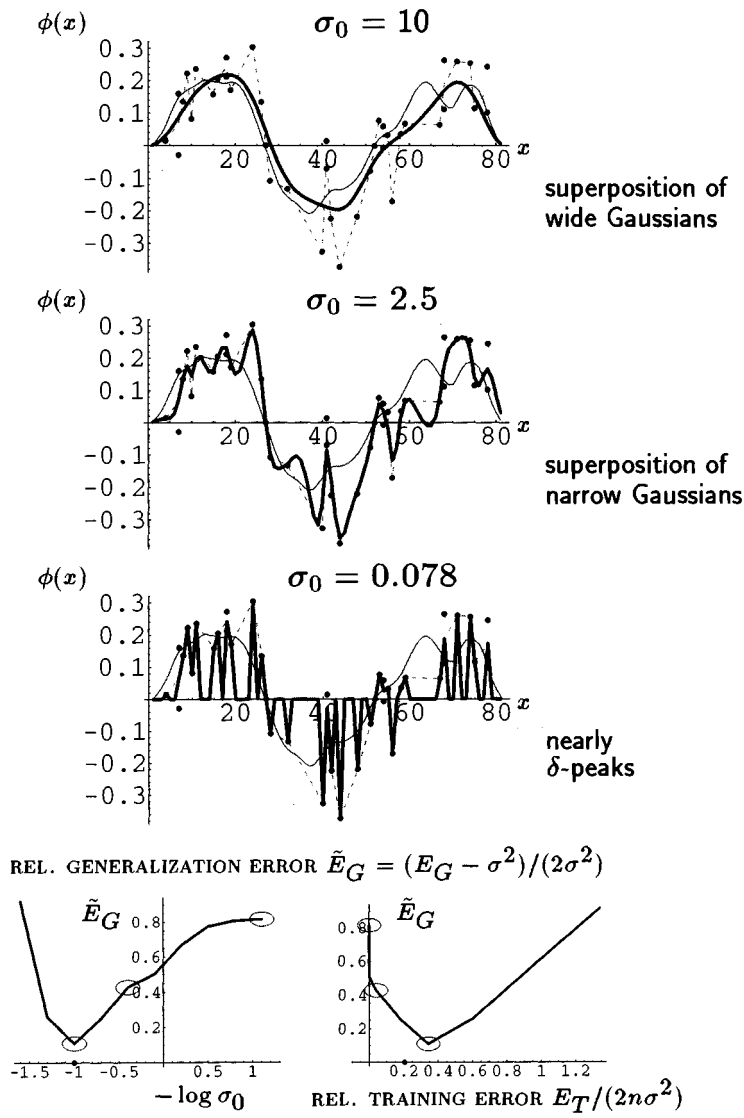


Figure 3.18: AIR 8: Regression with a truncated Radial Basis Function smoothness prior like in Fig. 3.17 but for varying σ_0 and fixed regularization factor $\lambda = 1$. As likelihood $p(y|x,\phi)$ a Gaussian likelihood has been chosen with standard deviation $\sigma = 1.8$ and a regression function $\phi_{\text{true}}(x)$ sampled according to the above prior. Training error and generalization error defined as for Fig. 3.16.

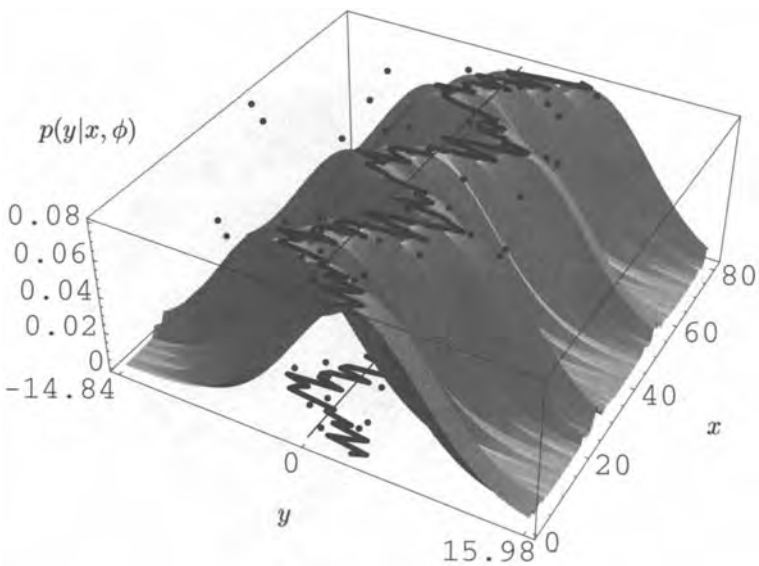


Figure 3.19: AIR 9: Approximately periodic Gaussian likelihood with negative Laplacian. Gaussian likelihood $p(y|x,\phi)$ of the form (2.11) with $\sigma = 5.0$ and regression function $\phi(x)$ (thick line) sampled from a Gaussian prior process (2.180) with $t_0(x) \equiv 0$ and inverse covariance $\mathbf{K}_0 = -0.9\Delta_\theta - 0.1\Delta$, as in (3.145), and $\theta = 27$. The inverse covariance $-\Delta_\theta$ defined in (3.141) implements approximate invariance under a finite translation.

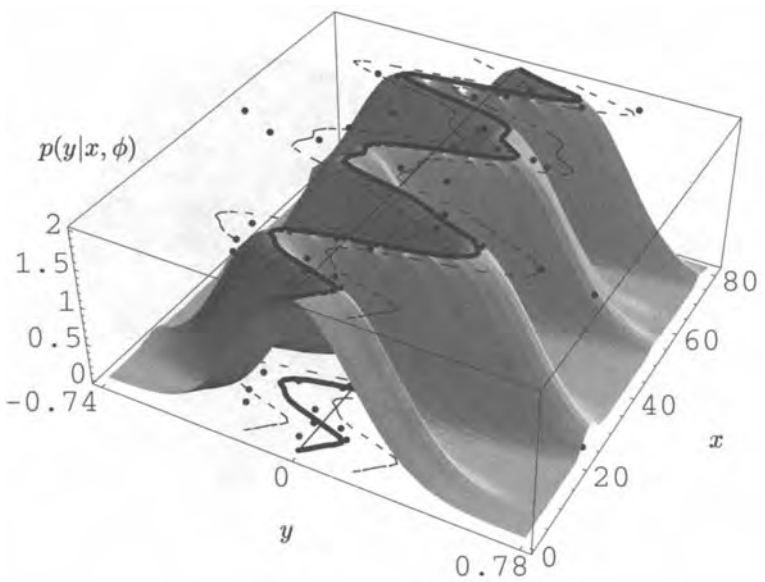


Figure 3.20: AIR 10: Approximately periodic Gaussian likelihood with RBF prior. Like Fig. 3.19 but with inverse covariance $\mathbf{K}_0 = -0.9\Delta_\theta + 0.1\sum_{k=0}^{3}\frac{1}{k!}\left(-\frac{\sigma_0^2\Delta}{2}\right)^k$, being a weighted sum of $-\Delta_\theta$, as defined in (3.141), and a truncated RBF prior as in (3.285) for $\theta = 27$, $\sigma_0 = 10$, $\sigma = 0.2$.

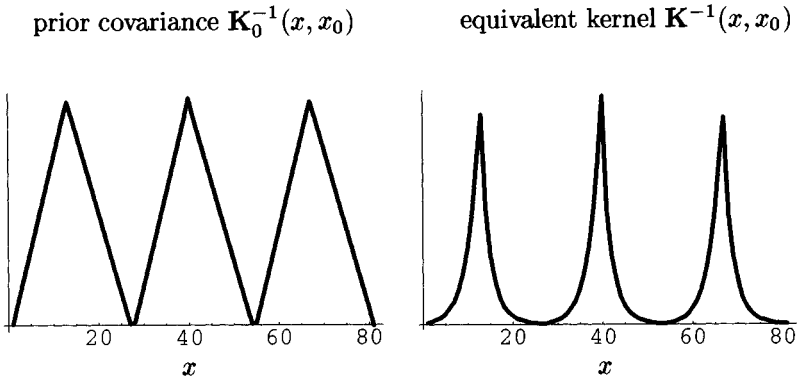


Figure 3.21: AIR 11: Approximate periodicity and smoothness 1: Prior covariance and equivalent kernel (like in Fig. 3.14) but for the prior of Fig. 3.19 being of the form (3.145) which implements approximate periodicity and smoothness. The equivalent kernel is based on the training data shown in Fig. 3.19.

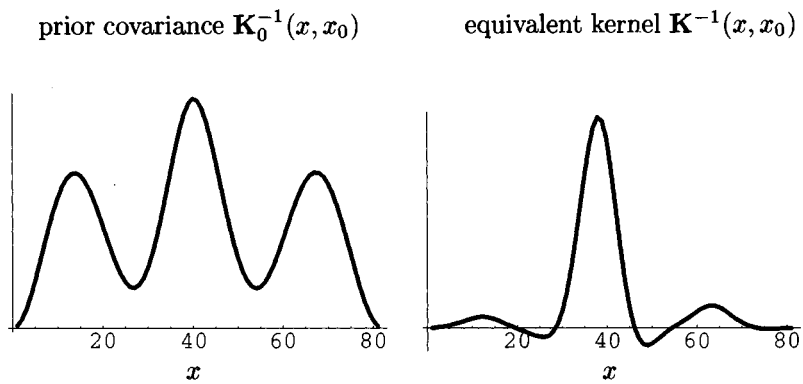


Figure 3.22: AIR 12: Approximate periodicity and smoothness 2: Like Fig. 3.21 but for the prior of Fig. 3.20 implementing approximate periodicity and smoothness using a truncated RBF prior as in (3.285). The equivalent kernel is based on the training data shown in Fig. 3.20.

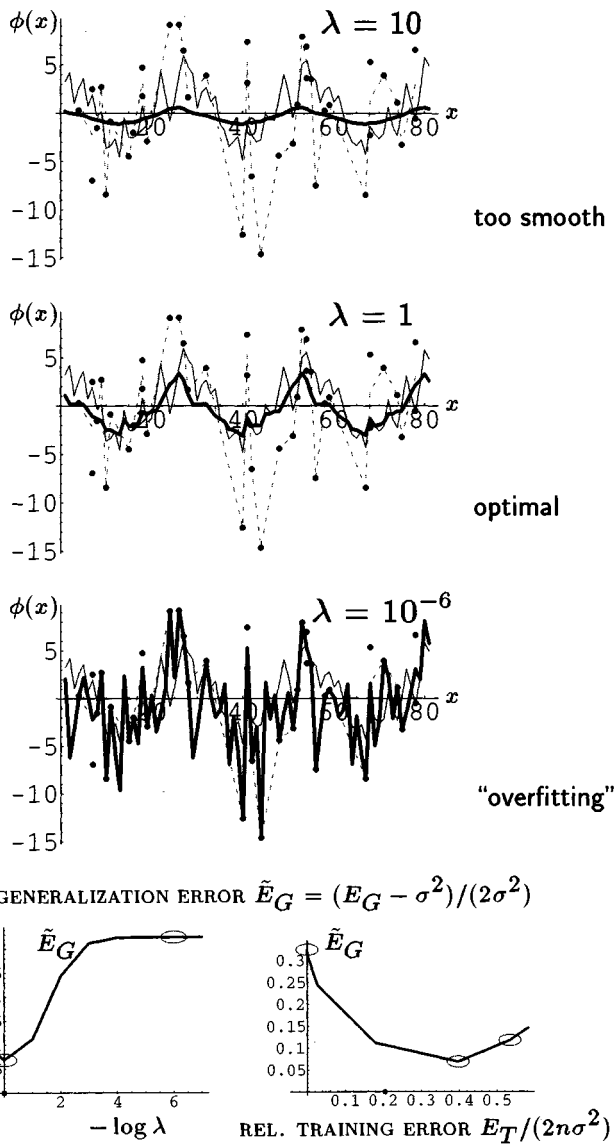


Figure 3.23: AIR 13: Approximately periodic regression 1. Regression with a prior implementing smoothness and approximate periodicity like in Fig. 3.19 with inverse covariance $\mathbf{K}_0 = -\lambda(0.9\Delta_\theta + 0.1\Delta)$, as in (3.145), with $\theta = 27$ and for varying λ. As likelihood $p(y|x,\phi)$ a Gaussian likelihood has been chosen with standard deviation $\sigma = 5.0$ and a regression function $\phi_{\text{true}}(x)$ sampled according to the above prior. Training error and generalization error defined as in Fig. 3.16.

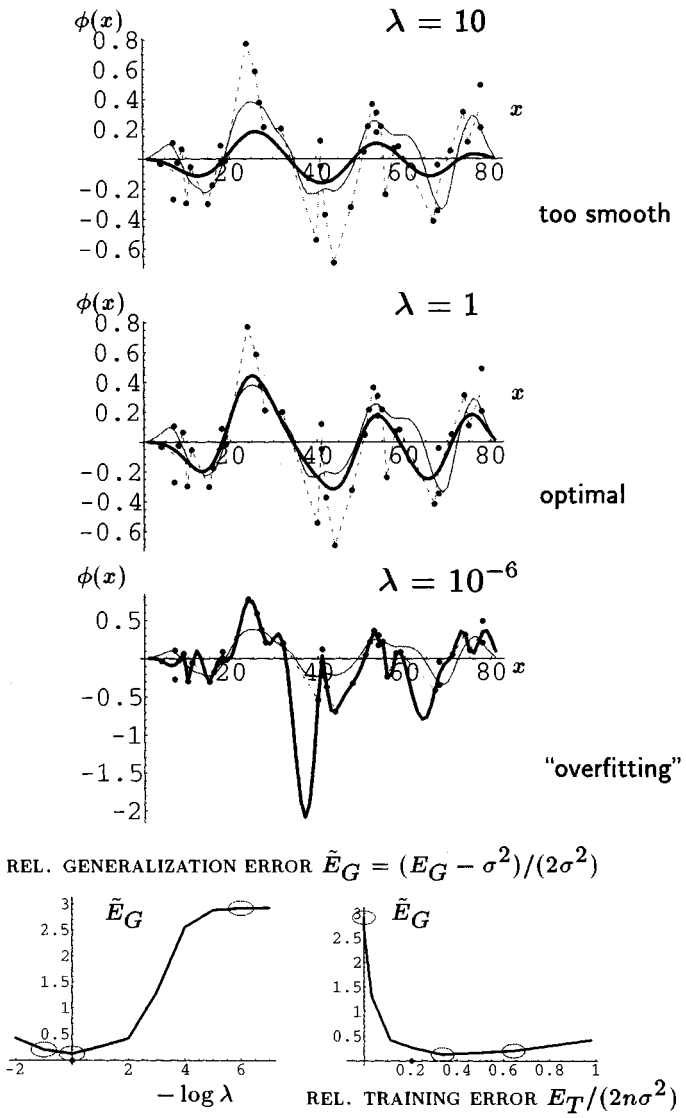


Figure 3.24: AIR 14: Approximately periodic regression 2. Reconstruction of approximately periodic functions like in Fig. 3.23 but with a truncated RBF smoothness prior $K_0 = \lambda \sum_{k=0}^{3} \frac{1}{k!} \left(-\sigma_0^2 \Delta/2\right)^k$ as in (3.285) with $\sigma_0 = 10$ replacing the negative Laplacian.

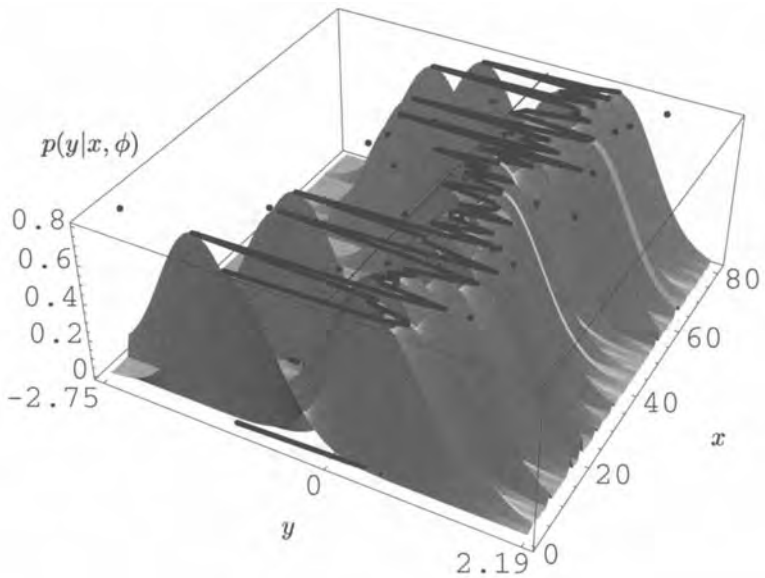


Figure 3.25: AIR 15: Approximately fractal Gaussian likelihood. Likelihood $p(y|x,\phi)$ of the form (2.11) with $\sigma = 0.5$ and regression function $\phi(x)$ (thick line) sampled from a Gaussian prior process (2.180) with $t_0(x) \equiv 0$ and inverse covariance $\mathbf{K}_0 = (1-\gamma)\mathbf{K}_{F3} - \gamma\Delta$, $\gamma = 0.1$, with $\mathbf{K}_{F3}$ defined in (3.154). For the example $m = 3$ iterations of a multiple copy operation have been implemented, each placing at each iteration step one of $M = 2$ copies of the whole interval in its left and right third, scaled down by a factor of $1/s_k = 3$.

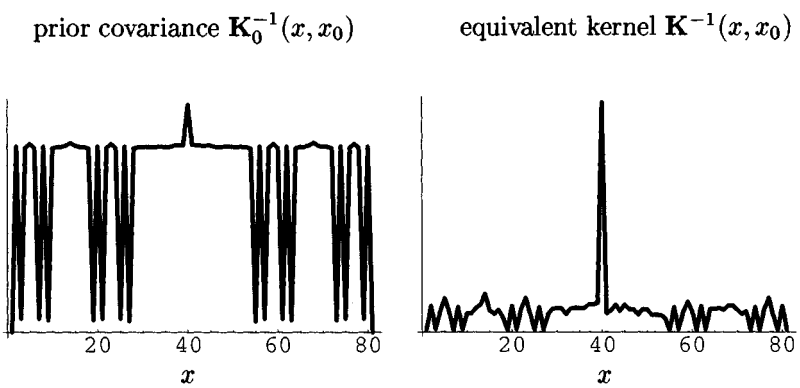


Figure 3.26: AIR 16: Approximate fractal: Prior covariance and equivalent kernel. Like Fig. 3.14 but for the prior of Fig. 3.25 possessing inverse covariance $\mathbf{K}_0 = 0.9\mathbf{K}_{F3} - 0.1\Delta$ with $\mathbf{K}_{F3}$ defined in (3.154) and implementing an approximate fractal structure. The equivalent kernel is based on the training data shown in Fig. 3.25.

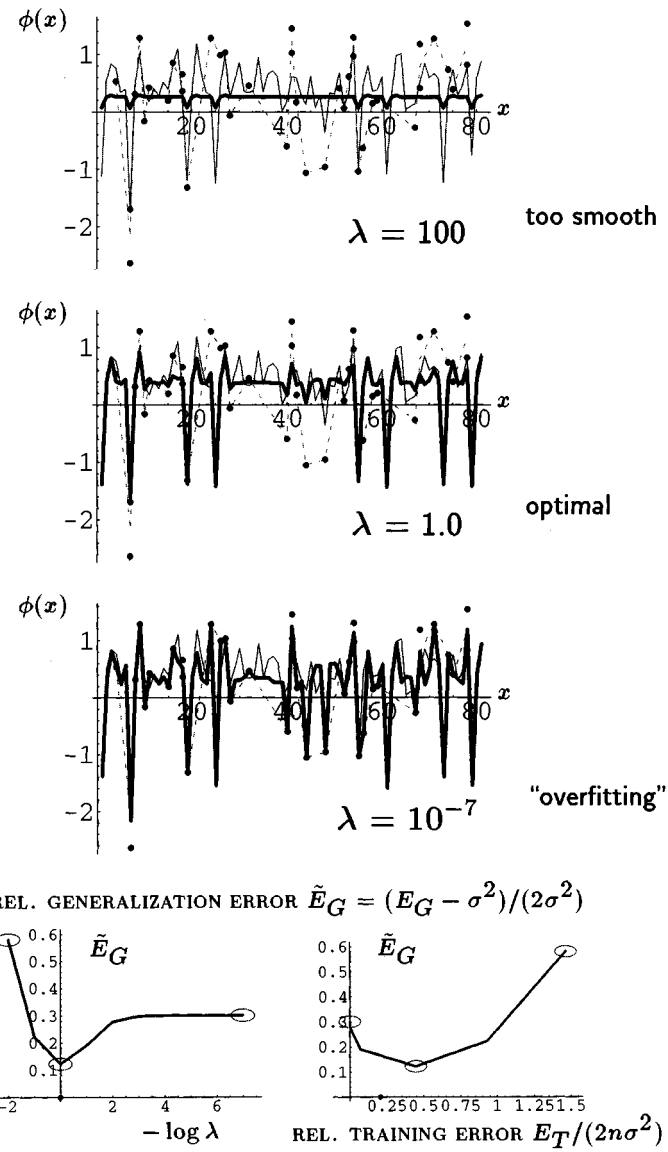


Figure 3.27: AIR 17; Approximately fractal regression. Reconstruction of an approximate fractal using an inverse prior covariance $\mathbf{K}_0 = \lambda[(1-\gamma)\mathbf{K}_{F3} - \gamma\Delta]$ for varying λ with $\mathbf{K}_{F3}$ as defined in (3.154) and $\gamma = 0.1$. The operator $\mathbf{K}_{F3}$ has been implemented with m =3 multiple copy operations each placing one of $M = 2$ copies of the whole interval in its left and right third, scaled down by a factor of $1/s_k = 3$. A Gaussian likelihood $p(y|x,\phi)$ has been chosen with standard deviation $\sigma = 0.5$ and an approximate fractal regression function $\phi_{\text{true}}(x)$ sampled according to the above prior.

3.9 Classification

In classification (or pattern recognition) tasks the visible variable y takes discrete values, representing a group, cluster or pattern label [33, 52, 106, 140, 209, 340, 379, 416]. Like in density estimation, the aim is to calculate the predictive probability (density) $p(y|x, D, D_0)$, defined in Eq. (2.64), for the group label y, given x, classification data $D = \{(x_i, y_i)|1 \leq i \leq n\} = (x_D, y_D)$ and *a priori* information D_0. Again, the ϕ–integral in Eq. (2.64) can be calculated in MAP or by using Monte Carlo methods.

If the predictive density is expressed in terms of the 'direct' likelihood $p(y|x, \phi)$ we will speak of a *direct learning model.* In that case, classification appears as a special case of the model (2.6) discussed in Section 2.1.1. Alternatively, classification can be formulated in terms of the 'indirect' likelihood $p(x|y, \phi)$. Such *indirect learning models*, often used for parametric approaches, are discussed below.

The direct model with a prior on the field ϕ corresponds to the density estimation error functional (3.157)

$$E_{\text{class.}}(\phi) = -\sum_i^n \ln P_{y_i}(x_i; \phi) + \frac{1}{2}\langle \phi - t\,|\mathbf{K}|\,\phi - t\rangle + \langle P(\phi)|\,\Lambda_X\,\rangle\,, \quad (3.287)$$

where y is discrete, so ϕ has components $\phi(x, y) = \phi_y(x)$ and the direct likelihood function P has components $p(y|x, \phi) = P(x, y) = P_y(x, \phi)$ with $\sum_y P_y(x, \phi) = 1$. In classification the scalar product thus corresponds to an integral over x and a summation over y,

$$\langle \phi - t\,|\mathbf{K}|\,\phi - t\rangle = \sum_{y,y'} \int dx\, dx' (\phi_y(x) - t_y(x))\mathbf{K}_{y,y'}(x, x')(\phi_{y'}(x') - t_{y'}(x')), \quad (3.288)$$

and, accordingly, $\langle P|\,\Lambda_X\,\rangle = \int dx\, \Lambda_X(x) \sum_y P_y(x)$.

For zero–one loss $l(x, y, a) = \delta_{y,a(x)}$ — a typical loss function for classification problems — the optimal decision (or *Bayes classifier*) is given by the mode of the predictive density (see Section 2.2.2), i.e.,

$$a(x) = \text{argmax}_y\, p(y|x, D, D_0). \quad (3.289)$$

In saddle point approximation $p(y|x, D, D_0) \approx p(y|x, \phi^*)$ where ϕ^* minimizing $E_{\text{class.}}(\phi)$ can be found by solving a stationarity equation analogous to Eq. (3.158).

As discussed for density estimation, the choice $\phi_y(x) = P_y(x)$ requires to take care of the non–negativity and normalization constraints explicitly. Similarly, for $\phi_y(x) = L_y(x)$ with $P_y(x) = e^{L_y(x)}$ a Lagrange multiplier has to be included to ensure normalization, but non–negativity is automatically fulfilled. Normalization is guaranteed by using unnormalized likelihoods $\phi_y(x) = z_y(x)$, $P_y(x) = z_y(x)/\sum_{y'} z_{y'}(x)$ (for which non–negativity has to be checked) or

shifted log–likelihoods $\phi_y(x) = g_y(x)$ with $g_y(x) = L_y(x) + \ln \sum_{y'} e^{L_{y'}(x)}$, i.e., $P_y(x) = e^{g_y(x)} / \sum_{y'} e^{g_{y'}(x)}$. In that case the nonlocal normalization terms are part of the likelihood and no Lagrange multiplier has to be used [521]. The resulting equation can be solved in data space (see Eq. (3.38)). The restriction of $\phi_y(x) = g_y(x)$ to linear functions $\phi_y(x) = w_y x + b_y$ yields log–linear models [333]. Recently a mean field theory for Gaussian process classification has been developed [374,377,378].

In contrast to direct models, indirect models are formulated in terms of the indirect likelihood $p(x|y,\phi)$ or $p(x|y,\xi)$, respectively, as we write for parametric models with parameters ξ. Indirect models are convenient if it is easier to give an appropriate parameterization for the probability (density) of finding x for given group y than for the probability of being in group y given x. For example, referring to expert knowledge it is typically easier to guess probabilities $p(x|y,\phi)$ for symptoms x given a disease y than to specify directly the probability $p(y|x,\phi)$ that a patient has disease y if he shows symptoms x. Thus, compared to a direct model, for an indirect model the roles of x and y as 'independent' and 'dependent' visible variables (as defined in Section 2.1.1) are exchanged. Nevertheless, in the indirect model, like in the direct model, one is finally interested in the ('direct') predictive density $p(y|x,\phi)$ to predict groups (e.g., diseases) y given findings (e.g., symptoms) x. A typical choice for a parametric indirect likelihood $p(x|y,\xi)$ is a Gaussian with mean, and possibly covariance, depending on the group y (leading to linear or quadratic discriminant analysis). A graphical representation of the indirect model is shown in Fig. 3.28.

For the indirect model we express the predictive density by indirect likelihoods

$$\begin{aligned} p(y|x,D,D_0) &= \int d\phi\, p(y|x,\phi)\, p(\phi|D,D_0) \\ &= \int d\phi\, \frac{p(x|y,\phi)\, p(y)}{\sum_y p(x|y,\phi)\, p(y|\phi)} p(\phi|D,D_0) \qquad (3.290) \\ &= \frac{p(x|y,D,D_0)\, p(y)}{\sum_y p(x|y,D,D_0)\, p(y)} \\ &= \frac{p(y) \int d\phi\, p(x|y,\phi)\, p(\phi|D,D_0)}{\sum_y p(y) \int d\phi\, p(x|y,\phi)\, p(\phi|D,D_0)}, \qquad (3.291) \end{aligned}$$

where for the indirect model $p(y|\phi) = p(y)$. While in the formulation of Eq. (3.290) a ϕ–dependent normalization $p(x|\phi)$ appears, ϕ is integrated out in the normalization factor $p(x|D,D_0)$ of Eq. (3.291). Thus, calculating the numerator of Eq. (3.291) in MAP, the MAP of the denominator can be obtained by summing over y weighted by $p(y)$. The posterior to be maximized,

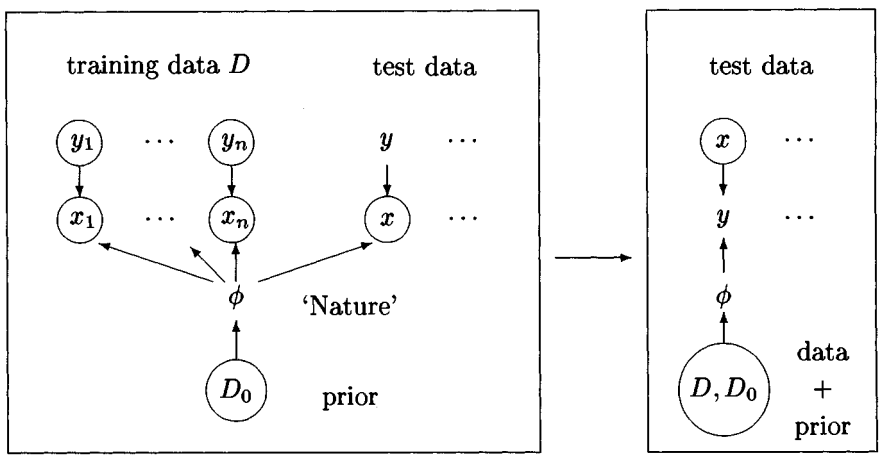


Figure 3.28: Graphical representation of an indirect learning model corresponding to the direct learning problem of Fig. 2.6. In contrast to the direct model, y represents the 'independent' visible variables and x the 'dependent' visible variables. The indirect model factorizes according to $p(x, x_D, y, y_D, \phi|D_0) = p(y)\, p(x|y, \phi)\, p(\phi|D_0) \prod_{i=1}^{n} p(y_i)p(x_i|y_i, \phi)$, as shown on the left. Given training data $D = (x_D, y_D) = \{(x_i, y_i)|1 \leq i \leq n\}$ and *a priori* information D_0, one is interested in the predictive probability or density $p(y|x, D.D_0)$ given in Eq. (3.291) for predicting outcome (e.g., group) y for a new x. The visible variables x, D, and D_0 conditioning the predictive density are indicated by circles.

reads for the indirect model

$$p(\phi|D, D_0) = \frac{p(D|\phi)p(\phi|D_0)}{p(D|D_0)} = \frac{p(x_D|y_D, \phi)p(\phi|D_0)}{p(x_D|y_D)}. \tag{3.292}$$

Notice that $p(y_D)$ cancels in the posterior (3.292), like $p(x_D)$ cancels in the posterior (2.60) for the direct model. However, according to the Bayesian inversion which is necessary to go from the indirect likelihood to the (direct) predictive density, Eq. (3.291) does depend on $p(y)$. In contrast, the predictive density (2.64) of the direct model, being conditional on x does not depend on $p(x)$. Finally we remark that indirect models can also be used for general density estimation problems if the indirect likelihood $p(x|y, \phi)$ is easier to parameterize than the direct $p(y|x, \phi)$.[6] An example of nonparametric classification using kernel methods can be found in Appendix C.3.3 and in Figure C.2.

Table 3.2 lists some special cases of density estimation. The last two lines

[6]The relation between direct and indirect models is also discussed in [291].

of the table, referring to inverse quantum mechanics, will be discussed in Section 7.

Likelihood $p(y\|x,\phi)$	Problem Type
$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\|y-\phi(x)\|^2}{2\sigma^2}\right]$	Gaussian regression
$[\phi(x)]^y e^{-\phi(x)}/y!\ {}^{y\in N,}_{\phi>0}$	Poisson regression
$p(y\|x,n;\phi)=\binom{n(x)}{y}[\phi(x)]^y[1-\phi(x)]^{n(x)-y}$ $(y\in N,\ y\le n\in N \text{ fixed},\ 0\le\phi\le 1)$	Binomial regression
$Ce^{-\|y-\phi(x)\|_\epsilon}$	ϵ-insensitive regression support vector regression
$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\|y-(\mathbf{F}\phi)(x)\|^2}{2\sigma^2}\right]$ (Gaussian, σ fixed)	classical inverse problems
$\sum_k \frac{1}{\sigma_k\sqrt{2\pi}}\exp\left[-\frac{[y-\phi_k(x)]^2}{2\sigma_k^2}\right]$ (mixture of Gaussians)	clustering
discrete y	classification
$p(y\|x,\phi)$ of general form	density estimation
$\mathrm{Trace}[\rho(\phi)\,P_{\hat{x}}(y)]$	Bayesian inverse quantum theory
$\langle\delta(y-x[x_{\mathrm{cl}},p_{\mathrm{cl}}])\rangle_{\mathrm{cl}}$	classical approximation of inverse quantum statistics

Table 3.2: Some special cases of density estimation. In inverse quantum theory ρ stands for the density operator of a quantum system, $\hat{x}$ for an observable, P_x for the projector into the space of eigenfunctions of $\hat{x}$ with eigenvalues y, and ϕ, for example, for a potential V to be reconstructed from data [23, 325, 336]. In the classical approximation x_{cl} and p_{cl} denote the classical coordinate and momentum. (The notations used in Chapter 7 differ slightly from the notations in this table, as in quantum mechanics the symbol x is commonly used to denote coordinates of particles.)

This page intentionally left blank

Chapter 4

Parameterizing likelihoods: Variational methods

4.1 General likelihood parameterizations

An approximate solution of the error minimization (or posterior maximization) problem is obtained by restricting the search (trial) space for $\phi(x, y)$ (or $\phi(x)$ in regression) to some parametric family $\phi(\xi)$. Functions ϕ which are in the considered search space are called *trial functions*. Solving a minimization problem in some restricted trial space is known as *variational approach* [56, 58, 80, 210, 240]. In this sense a MAP with a parametric model can be interpreted as a variational approach for a MAP for a nonparametric Bayesian problem. Clearly, minimal values obtained by minimization within a trial space can only be larger than or equal to the true minimal value, and from two variational approximations the one with smaller error is the better one.

Alternatively, the use of parameterized functions ϕ implements the *a priori* information that ϕ is known to have that specific parameterized form. (In cases where ϕ is only known to be approximately of a specific parameterized form, this should be implemented using a prior with a parameterized template and the parameters be treated as hyperparameters as in Section 5.)

Consider a field $\phi(x, y; \xi)$, parameterized by ξ, which determines the likelihood $p(y|x, \phi(\xi))$. Together with a range of allowed values for the components ξ_l of the parameter vector ξ this defines a possible trial space. Within a nonparametric Bayesian framework the hidden variable ϕ appears in two different roles: 1) It determines the likelihood and thus defines the training energy $-\sum_i \ln p(y_i|x_i, \phi)$ of the problem, and 2) it is used to formulate the prior $p(\phi)$ over likelihoods. For a parametric model, however, the prior may also be formulated directly in terms of the parameters $p(\xi)$. Accordingly there

are two variants to formulate priors for parameters.

1. Parametric prior as approximation of a nonparametric prior: Referring to an 'original' nonparametric model and interpreting a parametric model as a variational approximation, the 'original' prior $p(\phi)$ has to be translated into a prior $p(\xi) = p(\phi(\xi))$ for the parameters ξ.

2. Direct parametric prior: Interpreting a parametric model as the implementation of the *a priori* information that the likelihood has the given parametric structure, a prior $p(\xi)$ can be formulated directly in terms of ξ without referring to some 'original' nonparametric $p(\phi)$.

While the second method is the one which is commonly used for parametric models, we want in this section to study mainly the interpretation of parametric models as variational approximations of nonparametric models. In Section 4.2 we will discuss priors formulated directly in ξ and priors formulated in terms of a function $\psi(\xi)$ where ψ is not necessarily related to ϕ.

Interpreting parametric approaches as variational approximation, we start from a nonparametric posterior energy $E(\phi) = E(\phi|D_0, D)$, skipping again for the sake of simplicity the arguments D and D_0, and express ϕ by parameters ξ to obtain, for a Gaussian prior process on $\phi(x, y; \xi)$ with template $t(x, y)$ as in (3.157), a parameterized error functional

$$E(\phi(\xi)) = -\langle \ln P(\xi) | N \rangle + \frac{1}{2} \langle \phi(\xi) - t | \mathbf{K} | \phi(\xi) - t \rangle + \langle P(\xi) | \Lambda_X \rangle, \quad (4.1)$$

with ϕ depending on parameters ξ and $P(\xi) = P(\phi(\xi))$ the vector (function) with components $p(y|x, \phi(\xi))$. In the special case of Gaussian regression the positivity and normalization constraints are always fulfilled, so the Lagrange multiplier function Λ_X can be skipped and Eq. (4.1) becomes

$$E(\phi(\xi)) = \frac{1}{2} \langle \phi(\xi) - t_D | \mathbf{K}_D | \phi(\xi) - t_D \rangle + \frac{1}{2} \langle \phi(\xi) - t | \mathbf{K} | \phi(\xi) - t \rangle, \quad (4.2)$$

where the matrix elements include integrations over the x–variable only. Defining the matrix

$$\Phi'(l; x, y) = \frac{\partial \phi(x, y)}{\partial \xi_l}, \quad (4.3)$$

the stationarity equation for the functional (4.1) becomes

$$0 = \Phi' \mathbf{P}' \mathbf{P}^{-1} N - \Phi' \mathbf{K} (\phi - t) - \Phi' \mathbf{P}' \Lambda_X, \quad (4.4)$$

where $\mathbf{P}'$ denotes the Jacobi matrix (3.100). To have a convenient notation when solving for Λ_X we introduce

$$\mathbf{P}'_\xi = \Phi'(\xi) \mathbf{P}'(\phi), \quad (4.5)$$

i.e.,

$$\mathbf{P}'_\xi(l;x,y) = \frac{\partial P(x,y)}{\partial \xi_l} = \int dx'dy' \, \frac{\partial \phi(x',y')}{\partial \xi_l} \frac{\delta P(x,y)}{\delta \phi(x',y')}, \tag{4.6}$$

and

$$G_{\phi(\xi)} = \mathbf{P}'_\xi \mathbf{P}^{-1} N - \Phi' \mathbf{K}(\phi - t), \tag{4.7}$$

to obtain for Eq. (4.4)

$$\mathbf{P}'_\xi \Lambda_X = G_{\phi(\xi)}. \tag{4.8}$$

Eq. (4.4) may also be written in a form similar to Eq. (3.105)

$$\mathbf{K}_{\phi(\xi)}(\xi) = T_{\phi(\xi)} \tag{4.9}$$

with

$$T_{\phi(\xi)}(\xi) = \mathbf{P}'_\xi \mathbf{P}^{-1} N - \mathbf{P}'_\xi \Lambda_X + \Phi' \mathbf{K} t, \tag{4.10}$$

but with

$$\mathbf{K}_{\phi(\xi)}(\xi) = \Phi' \mathbf{K} \Phi(\xi), \tag{4.11}$$

being in general a nonlinear operator.

For a parameterization ξ, restricting the space of possible P, the matrix $\mathbf{P}'_\xi$ is not square and can not be inverted. Thus let $(\mathbf{P}'_\xi)^\#$ be the Moore–Penrose inverse of $\mathbf{P}'_\xi$,

$$(\mathbf{P}'_\xi)^\# \mathbf{P}'_\xi (\mathbf{P}'_\xi)^\# = (\mathbf{P}'_\xi)^\#, \quad \mathbf{P}'_\xi (\mathbf{P}'_\xi)^\# \mathbf{P}'_\xi = \mathbf{P}'_\xi, \tag{4.12}$$

and symmetric $(\mathbf{P}'_\xi)^\# \mathbf{P}'_\xi$ and $\mathbf{P}'_\xi (\mathbf{P}'_\xi)^\#$. A solution of Eq. (4.8) for Λ_X exists if

$$\mathbf{P}'_\xi (\mathbf{P}'_\xi)^\# G_{\phi(\xi)} = G_{\phi(\xi)}. \tag{4.13}$$

In that case the solution can be written as

$$\Lambda_X = (\mathbf{P}'_\xi)^\# G_{\phi(\xi)} + V_\Lambda - (\mathbf{P}'_\xi)^\# \mathbf{P}'_\xi V_\Lambda, \tag{4.14}$$

with arbitrary vector V_Λ and with

$$\Lambda^0_X = V_\Lambda - (\mathbf{P}'_\xi)^\# \mathbf{P}'_\xi V_\Lambda \tag{4.15}$$

in the right null space of $\mathbf{P}'_\xi$, representing a solution of

$$\mathbf{P}'_\xi \Lambda^0_X = 0. \tag{4.16}$$

Inserting, for $\Lambda_X(x) \neq 0$, Eq. (4.14) into the normalization condition $\Lambda_X = \mathbf{I}_X \mathbf{P} \Lambda_X$ gives

$$\Lambda_X = \mathbf{I}_X \mathbf{P} \left((\mathbf{P}'_\xi)^\# G_{\phi(\xi)} + V_\Lambda - (\mathbf{P}'_\xi)^\# \mathbf{P}'_\xi V_\Lambda \right). \tag{4.17}$$

Substituting back in Eq. (4.4) Λ_X is eliminated, yielding as stationarity equation

$$0 = \left(\mathbf{I} - \mathbf{P}'_\xi \mathbf{I}_X \mathbf{P} (\mathbf{P}'_\xi)^\# \right) G_{\phi(\xi)} - \mathbf{P}'_\xi \mathbf{I}_X \mathbf{P} \left(V_\Lambda - (\mathbf{P}'_\xi)^\# \mathbf{P}'_\xi V_\Lambda \right), \tag{4.18}$$

where $G_{\phi(\xi)}$ has to fulfill Eq. (4.13).

4.2 Gaussian priors for likelihood parameters

In Section 4.1 we have derived the prior $p(\phi(\xi))$ on ξ from a nonparametric prior $p(\phi)$ by parameterizing the function $\phi(x,y;\xi)$ of x and y. Instead of a prior for a function $\phi(x,y;\xi)$ a prior can also be formulated directly in terms of the parameters ξ or in terms of a function of the parameters $\psi(\xi)$, independent of x and y. For instance, a Gaussian prior in $\psi(\xi) = W_\psi \xi$ being a linear function of ξ, results in a prior which is also Gaussian in the parameters ξ, corresponding to the regularization term (prior energy)

$$\frac{1}{2}\langle \xi \,|W_\psi^T \mathbf{K}_\psi W_\psi|\, \xi\rangle = \frac{1}{2}\langle \xi \,|\mathbf{K}_\xi|\, \xi\rangle\,, \tag{4.19}$$

where $\mathbf{K}_\xi = W_\psi^T \mathbf{K}_\psi W_\psi$ is not an operator in a space of functions $\phi(x,y)$ but a matrix in the space of parameters ξ. The notation $\langle \xi\,|\mathbf{K}_\xi|\,\xi\rangle$ thus stands for a sum over the components ξ_l of the parameter vector ξ, $\sum_{ll'} \xi_l \mathbf{K}_{\xi,ll'} \xi_{l'}$. The results of Section 4.1 apply to this case provided the following replacement is made

$$\Phi' \mathbf{K} \phi \to \mathbf{K}_\xi \xi. \tag{4.20}$$

Similarly, a nonlinear ψ requires the replacement

$$\Phi' \mathbf{K} \phi \to \Psi' \mathbf{K}_\psi \psi, \tag{4.21}$$

where

$$\Psi'(k,l) = \frac{\partial \psi_l(\xi)}{\partial \xi_k}. \tag{4.22}$$

Thus in the general case where a Gaussian (specific) prior in $\phi(\xi)$ and in $\psi(\xi)$ is given, we obtain

$$\begin{aligned} E(\phi(\xi),\psi(\xi)) \;=\; & -\langle \ln P(\xi)|\, N\rangle + \langle P(\xi)|\, \Lambda_X\rangle \\ & + \frac{1}{2}\langle \phi(\xi)\,|\mathbf{K}|\,\phi(\xi)\rangle + \frac{1}{2}\langle \psi(\xi)\,|\mathbf{K}_\psi|\,\psi(\xi)\rangle\,, \end{aligned} \tag{4.23}$$

or, including also non–zero template functions (means) t for ϕ and t_ψ for ψ as discussed in Section 3.7,

$$\begin{aligned} E(\phi(\xi),\psi(\xi)) \;=\; & -\langle \ln P(\xi)|\, N\rangle + \langle P(\xi)|\, \Lambda_X\rangle \\ & + \frac{1}{2}\langle \phi(\xi) - t\,|\mathbf{K}|\,\phi(\xi) - t\rangle \\ & + \frac{1}{2}\langle \psi(\xi) - t_\psi\,|\mathbf{K}_\psi|\,\psi(\xi) - t_\psi\rangle\,. \end{aligned} \tag{4.24}$$

The ϕ and ψ–terms in Eq. (4.24) can be interpreted as an energy functional corresponding to a density $p(\xi|t,\mathbf{K},t_\psi,\mathbf{K}_\psi)$ or $p(t_\psi|\xi,\mathbf{K}_\psi)\; p(\xi|t,\mathbf{K})$ with one of the two terms corresponding to a Gaussian likelihood with ξ–independent normalization.

Variable	Error	Stationarity equation
$L(x,y)$	$E(L)$ Eqs. (3.157, 3.3)	$\mathbf{K}L = N - \mathbf{e}^{\mathbf{L}}\Lambda_X,$ $\Lambda_X = \mathbf{I}_X\,(N - \mathbf{K}L)$
$P(x,y)$	$E(P)$ Eqs. (3.157, 3.47)	$\mathbf{K}P = \mathbf{P}^{-1}N - \Lambda_X,$ $\Lambda_X = \mathbf{I}_X(N - \mathbf{P}\mathbf{K}P)$
$\phi = \sqrt{P}$	$E(\sqrt{P})$ Eqs. (3.157, 3.111)	$\mathbf{K}\phi = 2\Phi^{-1}N - 2\Phi\Lambda_X,$ $\Lambda_X = \mathbf{I}_X(N - \frac{1}{2}\Phi\mathbf{K}\phi)$
$\phi(x,y)$	$E(\phi)$ Eqs. (3.157)	$\mathbf{K}\phi = \mathbf{P}'\mathbf{P}^{-1}N - \mathbf{P}'\Lambda_X,$ $\Lambda_X = \mathbf{I}_X\left(N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi\right)$
ξ	$E(\phi(\xi))$ Eq. (4.1)	$\Phi'\mathbf{K}\phi = \mathbf{P}'_\xi\mathbf{P}^{-1}N - \mathbf{P}'_\xi\Lambda_X,$ $\Lambda_X = \mathbf{I}_X\mathbf{P}\left((\mathbf{P}'_\xi)^{\#}G_{\phi(\xi)} + \Lambda^0_X\right)$
ξ	$E(\phi(\xi),\psi(\xi))$ Eq. (4.24)	$\Phi'\mathbf{K}(\phi - t) + \Psi'\mathbf{K}_\psi(\psi - t_\psi)$ $= \mathbf{P}'_\xi\mathbf{P}^{-1}N - \mathbf{P}'_\xi\Lambda_X,$ $\Lambda_X = \mathbf{I}_X\mathbf{P}\left((\mathbf{P}'_\xi)^{\#}G_{\phi,\psi} + \Lambda^0_X\right)$

Table 4.1: Summary of stationarity equations. For notations, conditions, and comments see Sections 3.1.1, 3.2.1, 3.5.2, 3.5.1, 4.1, and 4.2.

The stationarity equation becomes

$$\begin{aligned} 0 &= \mathbf{P}'_\xi\mathbf{P}^{-1}N - \Phi'\mathbf{K}(\phi - t) - \Psi'\mathbf{K}_\psi(\psi - t_\psi) - \mathbf{P}'_\xi\Lambda_X \qquad (4.25)\\ &= G_{\phi,\psi} - \mathbf{P}'_\xi\Lambda_X, \qquad (4.26)\end{aligned}$$

which defines $G_{\phi,\psi}$, and for $\Lambda_X \neq 0$

$$\Lambda_X = \mathbf{I}_X\mathbf{P}\left((\mathbf{P}'_\xi)^{\#}G_{\phi,\psi} + \Lambda^0_X\right), \qquad (4.27)$$

for $\mathbf{P}'_\xi\Lambda^0_X = 0$.

4.3 Linear trial spaces

Solving a density estimation problem numerically, the function ϕ has to be discretized in some basis. This is done by expanding ϕ in a basis $B_l(x,y)$

(not necessarily orthonormal) and, choosing some $l_{\max}$, truncating the sum to terms with $l \leq l_{\max}$,

$$\phi = \sum_{l=1}^{\infty} c_l B_l \rightarrow \phi = \sum_{l=1}^{l_{\max}} c_l B_l, \tag{4.28}$$

where the numbers c_l represent the parameters ξ_l. This Ritz method corresponds to a finite linear trial space and is equivalent to solving a projected stationarity equation. Using a discretization (linear parameterization) of the form (4.28), the functional (3.99) becomes

$$E_{\text{Ritz}} = -\langle \ln P(\phi) | N \rangle + \frac{1}{2} \sum_{kl} c_k c_l \langle B_k | \mathbf{K} | B_l \rangle + \langle P(\phi) | \Lambda_X \rangle . \tag{4.29}$$

Non–zero templates $t(x, y)$ can be included by expanding them in the same truncated basis as ϕ

$$t = \sum_{l=1}^{\infty} \tilde{t}_l B_l \rightarrow t = \sum_{l=1}^{l_{\max}} \tilde{t}_l B_l, \tag{4.30}$$

and substituting $c_l \rightarrow c_l - \tilde{t}_l$ in Eq. (4.29). Solving for the coefficients c_l, $l \leq l_{\max}$ to minimize the error, results according to Eq. (4.4) and

$$\Phi'(l; x, y) = B_l(x, y) \tag{4.31}$$

in

$$0 = \langle B_l | \mathbf{P}' \mathbf{P}^{-1} N \rangle - \sum_k (c_k - \tilde{t}_k) \langle B_l | \mathbf{K} | B_k \rangle - \langle B_l | \mathbf{P}' \Lambda_X \rangle , \ \forall l \leq l_{\max}, \tag{4.32}$$

corresponding to the $l_{\max}$–dimensional equation for the vector c

$$\mathbf{K}_B \, c = N_B(c) - \Lambda_B(c) + \mathbf{K}_B \, \tilde{t}, \tag{4.33}$$

where

$$\mathbf{K}_B(l, k) = \langle B_l | \mathbf{K} | B_k \rangle , \tag{4.34}$$

$$N_B(c)(l) = \langle B_l | \mathbf{P}'(\phi(c)) \, \mathbf{P}^{-1}(\phi(c)) \, N \rangle , \tag{4.35}$$

$$\Lambda_B(c)(l) = \langle B_l | \mathbf{P}'(\phi(c)) \, \Lambda_X \rangle , \tag{4.36}$$

and c and $\tilde{t}$ denote the vectors with components $c(l) = c_l$ or $\tilde{t}(l) = \tilde{t}_l$, respectively. Thus for an orthonormal basis B_l, Eq. (4.33) corresponds to Eq. (3.101) projected into the trial space by the projector $\sum_l B_l \, B_l^T$.

Generalized linear models use linear approaches

$$\phi(z) = \sum_{l=0}^{m} c_l B_l = c_0 + \sum_{l=1}^{m} c_l z_l, \tag{4.37}$$

where $B_0 = 1$, $B_l = z_l$, and $z = (x, y)$, or, for regression, $z = x$, respectively [183, 333, 497]. Linear models for regression will be treated in Section 4.4. Dependent y– or independent x–variables which are not continuous but class variables (also called categorical, qualitative, discrete, or nominal variables) can be included by introducing zero–one dummy variables, like δ_{x1}, δ_{x2}, $\cdots$, δ_{ym} for an x–variable which can take the discrete values $1, \cdots, m$, or δ_{y1}, δ_{y2}, $\cdots$, δ_{ym} for an analogous discrete y. For both, x and y, discrete Eq. (4.37) becomes

$$\phi(x, y) = c_0 + \sum_{x'} c_{xx'} \delta_{xx'} + \sum_{y'} c_{yy'} \delta_{yy'}. \tag{4.38}$$

Clearly, continuous and discrete variables can be combined in one model equation.

Interactions, i.e., terms proportional to products of z–components like $c_{kl} z_k z_l$, can be included. Including all possible interactions corresponds to a multidimensional Taylor expansion of the function $\phi(z)$. Models with different linear expressions in x for different values of y can be expressed as linear models with interactions by introducing dummy variables. For example, for a classification model with m classes

$$\begin{aligned} \phi(x, y) &= \phi(x, (\delta_{y1}, \delta_{y2}, \cdots, \delta_{ym})) \\ &= \sum_{y'=1}^{m} \delta_{yy'} \left(c_{0y'} + \sum_l c_{ly'} x_l \right) \\ &= c_{0y} + \sum_l c_{ly} x_l, \end{aligned} \tag{4.39}$$

with y–dependent coefficients c_{ly}, which form the parameter vector $\xi = c$, and $\delta_{yy'}$ being zero–one dummy variables for the m classes so that y can be represented by the dummy vector δ_{y1}, δ_{y2}, $\cdots$, δ_{yM}.

In the special case of classification with a binary y,

$$p(y|x, c) = \frac{e^{-E_0(x,c)}}{e^{-E_0(x,c)} + e^{-E_1(x,c)}} = \frac{1}{1 + e^{\Delta E(x,c)}} \tag{4.40}$$

where, choosing zero or one as the two possible values of y, $E_0(x, c) = E(y = 0|x, c)$, $E_1(x, c) = E(y = 1|x, c)$, and $\Delta E(x, c) = E_0(x, c) - E_1(x, c)$.

In *binary logistic regression* [5] the energy difference (also known as *logit*)

$$\phi = \Delta E = \ln \frac{p_1}{p_0} = \ln \frac{p_1}{1 - p_1}, \tag{4.41}$$

writing $p_i(x, c) = p(y = i|x, c)$, is modeled as a linear funtion in the vector x

$$\phi(x) = \Delta E(x, c) = c_0 + \sum_{l=1}^{m} c_l x^l. \tag{4.42}$$

Similarly, for classification problems with more than two groups

$$p(y_k|x,c) = \frac{e^{-E_k(x,c)}}{\sum_k e^{-E_k(x,c)}}, \tag{4.43}$$

one can use *multinomial logistic regression*, which works with linear expressions in x for the energies

$$\phi(x,y_k) = E(y_k|x,c) = c_{0k} + \sum_{l=1}^{m} c_{lk}x^l, \tag{4.44}$$

i.e., $\phi_k(x) = E_k$ for $E_k(x,c) = E(y{=}k|x,c)$. One of the energy levels E_k has to be fixed to get a unique solution, because the likelihood (4.43) is invariant when adding a constant to the energy according to the explicit implementation of the normalization condition (cf. Section 3.1).

If there exists an ordinal relation between the groups, i.e., if there is a meaningful relation $y_1 < y_2 < \cdots < y_m$, then one can approximate the cumulative likelihoods $p_{\text{cum}}(y \geq k|x,c)$ instead of the $p(y = k|x,c)$. Using m different linear equations for the m cumulative likelihoods is known as *cumulative logistic regression*, using just m different constants c_{0k} but a common factor c_l for the x_l is known as *ordinal logistic regression*.

If the functions $B_l(z)$ are parameterized this leads to the mixture models for ϕ discussed in Section 4.5, if one allows lower dimensional nonparametric estimates for the B_l, this leads to additive models discussed in Section 4.6.

4.4 Linear regression

In this Section we take a closer look at linear regression. It goes back to Legendre and Gauss and is probably one of the most used models in statistics. As we will see, it already contains the structure of the general nonparametric regression model which we have discussed in Section 3.8.

Standard linear regression: To use the common notation for linear regression we will write $\hat{y}(x)$ for the regression function $\phi(x)$ being our approximation for $y(x)$. In linear regression $\hat{y}(x) = \phi(x)$ is restricted to a linear model

$$\hat{y}(x,c) = c_0 + \sum_{k=1}^{m-1} c_k x_k = \sum_{k=0}^{m-1} c_k x_k, \tag{4.45}$$

where x_k is the kth component of an m–dimensional vector x, and we introduced $x_0 = 1$ to be able to include the constant c_0 in the sum. As already mentioned in the last Section it is straightforward to include interaction terms

like $x_k x_l$ or $x_{k_1} x_{k_2} x_{k_3} \cdots$. Evaluated at the training points x_i this linear regression model reads

$$\hat{y}_i = \hat{y}(x_i, c) = c_0 + \sum_{k=1}^{m-1} c_k x_{ik} = \sum_{k=0}^{m-1} c_k x_{ik} = \sum_{k=0}^{m-1} \mathbf{X}_{ik} c_k, \tag{4.46}$$

with x_{ik} being the kth component of the ith training vector x_i, all $x_{i0} = 1$, and $n \times m$ *design matrix* $\mathbf{X}$ with entries $\mathbf{X}_{ik} = x_{ik}$, $1 \leq i \leq n$, $0 \leq k \leq m-1$. Introducing the n–dimensional vector $\hat{y}$ of approximations $\hat{y}_i$, and the m–dimensional coefficient vector c with components c_k, Eq. (4.46) reads in matrix notation

$$\hat{y} = \mathbf{X}c. \tag{4.47}$$

Assuming for y a Gaussian likelihood model with mean $\hat{y}(x, c)$ (with the unknown, true c to be reconstructed)

$$p(y|x, c) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|y - \hat{y}(x,c)|}{2\sigma^2}}, \tag{4.48}$$

the training energy $E(c) = -\langle \ln P(\hat{y})|\, N \rangle$ in (4.29) becomes the mean squared error

$$E(c) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \sum_{k=0}^{m-1} x_{ik} c_k)^2. \tag{4.49}$$

The Gaussian model (4.48) is also often written in the form

$$y(x) = \hat{y}(x, c) + e(x), \tag{4.50}$$

where the error $e(x)$ is a Gaussian random variable with zero mean, independent between different x and with x–independent variance σ^2. The realizations $e_i = y_i - \hat{y}_i$ of the error at the training data points are known as *residuals*. Introducing besides $\hat{y}$ also the two n–dimensional vectors y of training data y_i and e of residuals e_i, we thus have on the data points the matrix equation

$$y = \hat{y} + e. \tag{4.51}$$

If the number of (linearly independent) data is not smaller than the number m of unknown coefficients, then no prior is necessary to get a unique solution for c. Indeed, from (4.49) follows directly as stationarity equation for c, (also called *normal equation*)

$$0 = \frac{\partial E(c)}{\partial c_k} = \sum_{i=1}^{n} x_{ik} \left(y_i - \sum_{l=0}^{m-1} x_{il} c_l \right), \tag{4.52}$$

i.e., in matrix notation

$$\mathbf{X}^T y - \mathbf{X}^T \mathbf{X} c = 0. \tag{4.53}$$

The stationarity Eq. (4.53) is solved by

$$c = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{T}y, \tag{4.54}$$

where $\mathbf{X}^T\mathbf{X}$ is an $m \times m$–matrix. (Practical methods for solving Eq. (4.53), based, for example, on Householder transformations, can be found in [350].) The approximation $\hat{y}$ corresponding to the solution c of (4.54) is

$$\hat{y} = \mathbf{X}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}y. \tag{4.55}$$

The approximation $\hat{y}$ is extended to non–training data points by inserting the corresponding x, leaving c constant, i.e., $\hat{y}(x) = x^T c$. The matrix

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T} \tag{4.56}$$

mapping analogous to the influence matrix (3.206) in Section 3.8 the data vector y to its approximation $\hat{y}$,

$$\hat{y} = \mathbf{A}y \tag{4.57}$$

is in this case also known as projection matrix, as, due to the absence of a prior term, here indeed $\mathbf{A}^2 = \mathbf{A}$. (This means for example that replacing the data y by their approximation $\hat{y}$ does not change that (iterated) approximation $\hat{y}$. This is not true in the presence of prior terms which would in that case enter the iterated solution twice.)

It is also not difficult to calculate the Hessian H_{kl}. The Hessian is, according to Section 2.1.6, Eq. (2.89), the inverse covariance in second order, which is already exact for Gaussians, and allows the calculation of error bars for c, also called standard errors. (A small standard error for the coeffients c means that their value does not depend very much on the random choice of the training data.) Indeed, we obtain

$$H_{kl} = \frac{E(c)}{\partial c_k \partial c_l} = \sigma^{-2}\mathbf{X}^{T}\mathbf{X}, \tag{4.58}$$

which means, similarly to Eqs. (2.92, 2.93), that the covariance between two components of c is

$$\mathrm{cov}(c)_{kl} = \mathrm{cov}(c_k, c_l) = \langle c_k c_l\rangle - \langle c_k\rangle\langle c_l\rangle = H^{-1}_{kl} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\sigma^2, \tag{4.59}$$

which can also be verified directly

$$\mathrm{cov}(c) = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathrm{cov}(e)\mathbf{X}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\sigma^2, \tag{4.60}$$

as for independent, identically distributed (i.i.d.) errors $\mathrm{cov}(e) = \sigma^2\mathbf{I}$. This means in particular that the standard error for coefficient c_k is

$$\sigma(c_k) = \sqrt{\langle c_k^2\rangle - \langle c_k\rangle^2} = H_{kk}^{-1/2}. \tag{4.61}$$

Eqs. (4.52, 4.53) can be expressed in terms of the residuals as

$$\sum_{i=1}^{n} x_{ik} e_i = 0, \quad \text{or} \quad \mathbf{X}^T e = 0. \tag{4.62}$$

In particular, in models with a constant c_0, the zero component of (4.62) tells us, because of $x_0 = 1$, that the average residual is zero,

$$\sum_{i=1}^{n} e_i / n = \langle e \rangle = 0. \tag{4.63}$$

Linear regression with constraints: Linear regression models (4.45) without constant term c_0 do not have the necessary degree of freedom to always achieve the relation (4.63) for the average of the residuals on the data points. In such cases the average residual may differ from zero even if the true expectation of the random variable $e(x)$ is still assumed to be zero. As an example of linear regression with constraints, we study such models without c_0 and enforce $\langle e \rangle = 0$ (e.g., referring to an interpretation as risk minimization). Adding to the mean squared error (4.49) a Lagrange parameter term proportional to $\langle e \rangle$ yields

$$\begin{aligned} E(c) &= \frac{|e|^2}{2\sigma^2} + n\lambda \langle e \rangle = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{n} (y_i - \hat{y}_i) \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{m} x_{ik} c_k \right)^2 + \lambda \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{m} x_{ik} c_k \right). \end{aligned} \tag{4.64}$$

To find the c with minimal energy we set the partial derivatives with respect to the c_k to zero

$$0 = \frac{\partial E(c)}{\partial c_k} = \sum_{i=1}^{n} x_{ik} \left(y_i - \sum_{l=1}^{m} x_{il} c_l \right) + \lambda \sum_{i=1}^{n} x_{ik}, \tag{4.65}$$

or, introducing the n–dimensional vector 1_n with all components equal to 1,

$$0 = \mathbf{X}^T y - \mathbf{X}^T \mathbf{X} c + \lambda \mathbf{X}^T 1_n, \tag{4.66}$$

which is solved by

$$c = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \left(y + \lambda 1_n \right). \tag{4.67}$$

To determine the Lagrange parameter λ we insert (4.67) into the constraint $\langle e \rangle = 0$,

$$0 = 1_n^T y - 1_n{}^T \mathbf{X} c = 1_n^T y - 1_n{}^T \mathbf{X} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \left(y + \lambda 1_n \right). \tag{4.68}$$

It follows, if $1_n^T \mathbf{A} 1_n \neq 0$ (there are cases where no solution exists),

$$\lambda = \frac{1_n^T \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) y}{1_n^T \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T 1_n} = \frac{1_n^T (\mathbf{I} - \mathbf{A})\, y}{1_n^T \mathbf{A} 1_n}, \tag{4.69}$$

to be inserted into (4.67), yielding

$$\hat{y} = \mathbf{X}c = \mathbf{A}\left(\mathbf{I} + 1_n \frac{1_n^T (\mathbf{I} - \mathbf{A})}{1_n^T \mathbf{A} 1_n}\right) y. \tag{4.70}$$

Linear regression with priors: Let us now consider the case where the number of data is smaller than the number of coefficients, i.e., $m > n$, or that the matrix $\mathbf{A}$ is ill-conditioned so small changes (like numerical errors) in y lead to large changes in the resulting approximation $\hat{y}$. In such cases additional *a priori* information is required. For example, preferring small coefficients one can choose a prior energy term like

$$E_0(c) = \frac{\lambda}{2} \sum_{k=0}^{m-1} c_k^2, \tag{4.71}$$

which, being quadratic, corresponds to a Gaussian prior $p(c) \propto e^{-E_0(c)}$. A general Gaussian prior on the coefficient vector c, with mean t_c and $m \times m$ inverse covariance $\mathbf{K}_0$, yields prior energy $\frac{1}{2}\langle c - t_0 | \mathbf{K}_0 | c - t_0 \rangle$ or, formulated in terms of $\mathbf{X}_0 c$ with some $m \times m$–prior design matrix $\mathbf{X}_0$,

$$E_0(c) = \frac{1}{2}\langle \mathbf{X}_0 c - t_0 | \mathbf{K}_0 | \mathbf{X}_0 c - t_0 \rangle, \tag{4.72}$$

like the prior term on $\psi(\xi)$ in (4.24). To combine data points y_i belonging to the same value of x_i (in case there is more than one) we can, like in Section 3.8, define the data vector t_D as in (3.176) or (3.183), containing for each x in the data the average of the corresponding y, and diagonal $\tilde{n} \times \tilde{n}$ data matrix $\mathbf{K}_D$ as in (3.175) or (3.181), containing on the diagonal the number of measurements at the corresponding x divided by σ^2. Again $\tilde{n} \leq n$ denotes the number of different x_i values in the data. We then work with the $\tilde{n}$–dimensional reduced approximation vector $\tilde{\hat{y}}$ and with the corresponding reduced $\tilde{n} \times m$ design matrix $\widetilde{\mathbf{X}}$ containing only distinct x_i values. If the values of all x_i are different, then $\tilde{n} = n$, $\tilde{\hat{y}} = \hat{y}$, and $\widetilde{\mathbf{X}} = \mathbf{X}$. Then the mean squared error (4.49) reads

$$\frac{1}{2\sigma^2}\langle \mathbf{X}c - y | \mathbf{X}c - y \rangle = \frac{1}{2}\left\langle \widetilde{\mathbf{X}}c - t_D \middle| \mathbf{K}_D \middle| \widetilde{\mathbf{X}}c - t_D \right\rangle + E_{\min}^D \tag{4.73}$$

with c–independent $E_{\min}^D$ as in (3.178, 3.179). Together with the prior (4.72), this leads us to the energy

$$E(c) = \frac{1}{2}\left\langle \widetilde{\mathbf{X}}c - t_D \middle| \mathbf{K}_D \middle| \widetilde{\mathbf{X}}c - t_D \right\rangle + \frac{1}{2}\langle \mathbf{X}_0 c - t_0 | \mathbf{K}_0 | \mathbf{X}_0 c - t_0 \rangle, \tag{4.74}$$

which possesses the same structure as the (discretized) energy (3.198) of the generalized regression problem (ignoring convergence issues), with the correspondencies:

$$\widetilde{\mathbf{X}}c = \widetilde{\hat{y}} \;\to\; \widetilde{\phi} = \mathbf{F}_D\phi, \quad c \to \phi, \quad \widetilde{\mathbf{X}} \to \mathbf{F}_D, \quad \mathbf{X}_0 \to \mathbf{F}_0. \tag{4.75}$$

The reason for this similarity is that from the nonparametric point of view (of functional analysis) functions are vectors and function values are vector components. In practice however, the dimensionality of the equations is typically extremely different in linear regression problems and nonparametric regression problems.

Using the correspondencies (4.75) we can read off the solution for (4.74) from Eq. (3.203)

$$\widetilde{\hat{y}} = \widetilde{\mathbf{X}}\,c = \widetilde{\mathbf{X}}\left(\widetilde{\mathbf{X}}^T\mathbf{K}_D\widetilde{\mathbf{X}} + \mathbf{F}_0^T\mathbf{K}_0\mathbf{F}_0\right)^{-1}\left(\widetilde{\mathbf{X}}^T\mathbf{K}_D t_D + \mathbf{X}_0^T\mathbf{K}_0 t_0\right). \tag{4.76}$$

The approximation $\widetilde{\hat{y}}$ is extended to $\hat{y}(x)$ for non–training data points x according to $\hat{y}(x) = x^T c$.

4.5 Mixture models

Generalizing linear models of the form (4.28), a function $\phi(z)$ can be approximated by a linear combination of component functions (mixture model)

$$\phi(z) = \sum c_l B_l(\xi_l, z), \tag{4.77}$$

but with a parameter vector ξ_l adaptable for each B_l and constants c_l (which can also be included into the vector ξ_l to be adapted). The functions $B_l(\xi_l, z)$ are often chosen to depend on one–dimensional combinations of vectors ξ_l and z. For example, choosing the ξ_l as vectors of the same dimension as z, they may depend on some distance $||\xi_l - z||$ ('local or distance approaches') or the projection of z in ξ_l–direction, i.e., $\sum_k \xi_{l,k} z_k$ ('projection approaches'). (For projection approaches see also Sections 4.6, 4.9, and 4.10). Increasing the flexibility of the parametric families $B_l(\xi_l, z)$ leads to the additive models discussed in Section 4.6.

A typical example are Radial Basis Functions (RBF) using a Gaussian $B_l(\xi_l, z)$ for which centers (and possibly covariances and number of components) can be adjusted. Other local approaches are k–nearest neighbors methods (kNN), learning vector quantizations (LVQ) [209, 269, 270], and its variants.

4.6 Additive models

Trial functions ϕ may be chosen as sum of simpler functions ϕ_l each depending only on part of the x– and y–variables. More precisely, we consider a set of

functions ϕ_l depending on projections $z_l = \mathbf{I}_l^{(z)} z$ of the vector $z = (x, y)$ of all x and y components. Thus $\mathbf{I}_l^{(z)}$ denotes a projector in the vector space of z (and not in the space of functions $\Phi(x, y)$). Hence, ϕ takes the form

$$\phi(z) = \sum_l \phi_l(z_l), \tag{4.78}$$

so only one–dimensional functions ϕ_l have to be determined. Restricting the functions ϕ_l to a parameterized function space yields a 'parameterized additive model'

$$\phi(z) = \sum_l \phi_l(\xi, z_l), \tag{4.79}$$

which has to be solved for the parameters ξ. The model can also be generalized to a model 'additive in parameters ξ_l'

$$\phi(z) = \sum_l \phi_l(\xi_l, x, y), \tag{4.80}$$

where the functions $\phi_l(\xi_l, x, y)$ are not restricted to one–dimensional functions depending only on projections z_l on the coordinate axes. If the parameters ξ_l determine the component functions ϕ_l completely, this yields just the mixture models of Section 4.5. Another example is projection pursuit, discussed in Section 4.9, where a parameter vector ξ_l of the same dimension as z defines a projection $\xi_l \cdot z$. In that case even for given ξ_l still the one–dimensional functions $\phi_l(\xi_l \cdot z)$ have to be determined.

An ansatz like (4.78) is made more flexible by including also 'interaction terms'

$$\phi(x, y) = \sum_l \phi_l(z_l) + \sum_{kl} \phi_{kl}(z_k, z_l) + \sum_{klm} \phi_{klm}(z_k, z_l, z_m) + \cdots. \tag{4.81}$$

The functions $\phi_{kl\ldots}(z_k, z_l, \cdots)$ can be chosen to depend on product terms like $z_{l,i} z_{k,j}$, or $z_{l,i} z_{k,j} z_{m,n}$, where $z_{l,i}$ denotes one–dimensional sub-variables of $z_l = \mathbf{I}_l^{(z)} z$.

In additive models in the narrower sense [201–203,473] z_l is a subset of x, y components, i.e., $z_l \subseteq \{x_i | 1 \leq i \leq d_x\} \cup \{y_j | 1 \leq j \leq d_y\}$, d_x denoting the dimension of x, d_y the dimension of y. In regression, for example, one takes usually the one–element subsets $z_l = \{x_l\}$ for $1 \leq l \leq d_x$. In more general schemes the projections of z do not have to be restricted to projections on the coordinates axes. In particular, the projections can be optimized too. For example, one–dimensional projections $\mathbf{I}_l^{(z)} z = w \cdot z$ with $z, w \in X \times Y$ (where $\cdot$ denotes a scalar product in the space of z variables) are used by *ridge approximation* schemes. They include for regression problems one–layer (and similarly multilayer) feedforward neural networks (see Section 4.10), projection pursuit regression (see Section 4.9), ridge regression [330,331], and hinge functions [63]. For a detailed discussion of the regression case see [162].

The stationarity equation for $E(\phi)$ becomes for the ansatz (4.78)

$$0 = \mathbf{P}_l' \mathbf{P}^{-1} N - \mathbf{K}(\phi - t) - \mathbf{P}_l' \Lambda_X, \tag{4.82}$$

with

$$\mathbf{P}_l'(z_l, z') = \frac{\delta P(z')}{\delta \phi_l(z_l)}. \tag{4.83}$$

Considering a density P, decomposed into components P_l, determined by the components ϕ_l,

$$P(z) = \sum_l P_l(\phi_l(z_l)), \tag{4.84}$$

the derivative (4.83) becomes

$$\mathbf{P}_l'(z_l, z_l') = \frac{\delta P_l(z_l')}{\delta \phi_l(z_l)}. \tag{4.85}$$

Specifying an additive prior energy

$$\frac{1}{2} \sum_{kl} \langle \phi_k - t_k \, | \mathbf{K}_{kl} | \, \phi_l - t_l \rangle, \tag{4.86}$$

the stationary conditions are coupled equations for the component functions ϕ_l which, because $\mathbf{P}$ is diagonal, only contain integrations over z_l–variables

$$0 = \frac{\delta P_l}{\delta \phi_l} \mathbf{P}^{-1} N - \sum_k \mathbf{K}_{lk}(\phi_k - t_k) - \frac{\delta P_l}{\delta \phi_l} \Lambda_X. \tag{4.87}$$

For the parameterized approach (4.79) one finds

$$0 = \mathbf{\Phi}_l' \mathbf{P}_l' \mathbf{P}^{-1} N - \mathbf{\Phi}_l' \mathbf{K}(\phi - t) - \mathbf{\Phi}_l' \mathbf{P}_l' \Lambda_X, \tag{4.88}$$

with

$$\mathbf{\Phi}_l'(k, z_l) = \frac{\partial \phi_l(z_l)}{\partial \xi_k}. \tag{4.89}$$

For the ansatz (4.80), the analogous $\mathbf{\Phi}_l'(k, z)$ is restricted to a subset of ξ_k.

4.7 Product ansatz

A product ansatz has the form

$$\phi(z) = \prod_l \phi_l(z_l), \tag{4.90}$$

where $z_l = \mathbf{I}_l^{(z)} z$ represents projections of the vector z consisting of all x and y components. The ansatz can be made more flexible by using a sum of products

$$\phi(z) = \sum_k \prod_l \phi_{k,l}(z_l). \tag{4.91}$$

The restriction of the trial space to product functions corresponds to the Hartree approximation in physics. (A straightforward extension of linear models is obtained by including an additional projection of the product functions. In a Hartree–Fock approximation, for example, product functions are antisymmetrized under coordinate exchange, i.e., projected into the space of antisymmetric functions.)

For additive $\mathbf{K} = \sum_l \mathbf{K}_l$ with $\mathbf{K}_l$ acting only on $\phi_l = \mathbf{I}_l \phi$, with some projector $\mathbf{I}_l$, $\mathbf{I}_l \, \mathbf{I}_{l'} = \delta_{l,l'}$, and factorizing template

$$t(z) = \prod_l \tilde{t}_l(z_l), \tag{4.92}$$

a quadratic regularization term becomes for the product ansatz (4.90),

$$\langle \phi - t \, | \mathbf{K} | \, \phi - t \rangle = \sum_l \langle \phi_l - \tilde{t}_l \, | \mathbf{K}_l | \, \phi_l - \tilde{t}_l \rangle \prod_{l' \neq l} \langle \phi_{l'} - \tilde{t}_{l'} \, | \, \phi_{l'} - \tilde{t}_{l'} \rangle \, . \tag{4.93}$$

Similarly, for $\mathbf{K} = \prod_l \mathbf{K}_l$ with a product structure with respect to ϕ_l

$$\langle \phi - t \, | \mathbf{K} | \, \phi - t \rangle = \prod_l \langle \phi_l - \tilde{t}_l \, | \mathbf{K}_l | \, \phi_l - \tilde{t}_l \rangle \, . \tag{4.94}$$

In both cases the prior term factorizes into lower dimensional contributions.

4.8 Decision trees

Decision trees [64] implement functions which are piecewise constant on rectangular areas parallel to the coordinate axes z_l. Such an approach can be written in tree structure with nodes only performing comparisons of the form $x < a$ or $x > a$ which allows a very effective hardware implementation. Such a piecewise constant approach can be written in the form

$$\phi(z) = \sum_l \xi_l \prod_k \Theta(z_{\xi_{lk}} - \xi_{0,lk}), \tag{4.95}$$

with step function Θ and $z_{\xi_{lk}}$ indicating the (l– and k–dependent) component ξ_{lk} of the vector z, which is compared with the reference value $\xi_{0,lk}$. There are effective constructive methods to build trees. Gradient–based optimization methods can be applied when replacing the step function by a sigmoid. In particular, decision trees correspond to neural networks at zero temperature, where sigmoids become step functions, and which are restricted to weight vectors in coordinate directions (see Section 4.10).

An overview of different variants of decision trees, together with a comparison with rule–based systems, neural networks (see Section 4.10); techniques from applied statistics like linear discriminants, projection pursuit (see Section 4.9), and local methods like k-nearest neighbors methods (kNN); Radial Basis Functions (RBF); or learning vector quantization (LVQ) is given in [340].

4.9 Projection pursuit

Projection pursuit models [108,110,138,227] are a generalization of additive models (4.78) (and a special case of models (4.80) additive in parameters) where the projections of $z = (x, y)$, defined by parameter vectors ξ_l of the same dimension as z, are also adapted

$$\phi(z) = \xi_0 + \sum_l \phi_l(\xi_{0,l} + \xi_l \cdot z). \tag{4.96}$$

For such a model one has to determine one–dimensional 'ridge' functions ϕ_l together with the projections defined by vectors ξ_l and the constants ξ_0, $\xi_{0,l}$. Adaptive projections may also be used for product approaches

$$\phi(z) = \prod_l \phi_l(\xi_{0,l} + \xi_l \cdot z). \tag{4.97}$$

Similarly, ϕ may be decomposed into functions depending on distances to adapted reference points (centers), giving models of the form

$$\phi(z) = \prod_l \phi_l(||\xi_l - z||), \tag{4.98}$$

which require the adaption of parameter vectors (centers) ξ_l and distance functions ϕ_l. For high dimensional spaces the number of centers necessary to cover a high dimensional space with fixed density grows exponentially. Furthermore, as the volume of a high dimensional sphere tends to be concentrated near its surface, the tails become more important in higher dimensions. Thus typically projection methods are better suited for high dimensional spaces than distance methods [456].

4.10 Neural networks

In projection pursuit–like techniques the one–dimensional 'ridge' functions ϕ_l (and the projections ξ_l and constants ξ_0, $\xi_{0,l}$) are adapted optimally. Similarly, neural networks also use ridge functions ϕ_l, but fixed to a sigmoidal form. The resulting lower flexibility following from fixing the ridge function is then compensated by iterating this parameterization. This leads to multilayer neural networks.

Multilayer neural networks have become a popular tool for regression and classification problems [9,10,24,52,81,141,168,206,209,273,279,343,344,358, 379,416,427,429,433,510,511]. One-layer neural networks (perceptrons[1] [344])

[1]Extensive literature exists on the perceptron from a statistical mechanics perspective based on the replica approach [143–145] as invented for spin glasses [128,338]. See also, for example, [375,376,511,527].

correspond to the parameterization

$$\phi(z) = \sigma\left(\sum_l w_l z_l - b\right) = \sigma(v), \tag{4.99}$$

with sigmoidal function σ, parameters $\xi = w$, projection $v = \sum_l w_l z_l - b$, and z_l single components of the variables x, y, i.e., $z_l = x_l$ for $1 \leq l \leq d_x$ and $z_l = y_l$ for $d_x + 1 \leq l \leq d_x + d_y$. (For neural networks with Lorentzians instead of sigmoids see [157].)

Typical choices for the sigmoid are $\sigma(v) = \tanh(\beta v)$ or $\sigma(v) = 1/(1 + e^{-2\beta v})$. The parameter β (inverse temperature) controls the sharpness of the step of the sigmoid. In particular, the sigmoid functions become a sharp step in the limit $\beta \to \infty$, i.e., at zero temperature. In principle the sigmoidal function σ may depend on further parameters which then — similar to projection pursuit discussed in Section 4.9 — also have to be included in the optimization process. The threshold or bias b can be treated as weight if an additional input component is included clamped to the value 1.

A linear combination of perceptrons

$$\phi(x, y) = b + \sum_l W_l\, \sigma\left(\sum_k w_{lk} z_k - b_k\right) \tag{4.100}$$

has the form of a projection pursuit approach (4.96) but with fixed $\phi_l(v) = W_l \sigma(v)$.

In multi–layer networks the parameterization (4.99) is cascaded,

$$z_{k,i} = \sigma\left(\sum_{l=1}^{m_{i-1}} w_{kl,i} z_{l,i-1} - b_{k,i})\right) = \sigma(v_{k,i}), \tag{4.101}$$

with $z_{k,i}$ representing the output of the kth node (neuron) in layer i and

$$v_{k,i} = \sum_{l=1}^{m_{i-1}} w_{kl,i} z_{l,i-1} - b_{k,i}, \tag{4.102}$$

being the input for that node. This yields, skipping the bias terms for simplicity,

$$\phi(z, w) = \tag{4.103}$$

$$\sigma\left(\sum_{l_{n-1}}^{m_{n-1}} w_{l_{n-1},n} \sigma\left(\sum_{l_{n-2}}^{m_{n-2}} w_{l_{n-1} l_{n-2},n-1} \cdots \sigma\left(\sum_{l_0}^{m_0} w_{l_1 l_0,1} z_{l_0,0}\right) \cdots\right)\right),$$

beginning with an input layer with $m_0 = d_x + d_y$ nodes $z_{l,0} = z_l$ (plus possibly nodes to implement the bias) and going over intermediate layers with m_i nodes $z_{l,i}$, $0 < i < n$, $1 \leq l \leq m_i$ to a single node output layer $z_n = \phi(x, y)$.

Neural nets are most commonly used in regression and classification to parameterize a function $\phi(x)$ to minimize a pure training error

$$E(w) = \sum_i (y_i - \phi(x_i, w))^2, \tag{4.104}$$

quadratic in ϕ and without further regularization (prior) terms. To avoid overfitting due to the missing prior terms, the necessary regularization has to be assured either

1. by using a neural network architecture which is restrictive enough, or
2. by starting from a reasonable initial guess for ϕ (like $\phi \equiv 0$) and stopping (or slowing down) the selected iterative training procedure before the network is able to develop its full flexibility. This technique is known as 'early stopping'.

In both cases the optimal architecture or algorithm can be determined, for example, by cross–validation [12, 89, 169, 356, 456, 471, 472, 505–507, 509] or bootstrap techniques [118–120, 193]. A third, Bayesian alternative consists of

3. averaging over ensembles of networks [363].

In all these cases regularization is implicit in the parameterization of the network. Alternatively, explicit regularization or prior terms can be added to the functional. For regression or classification this is, for example, done in *learning by hints* [2–4, 369] or *curvature–driven smoothing* with feedforward networks [51].

One may also remark that from a Frequentist point of view the quadratic functional (4.104) is not interpreted as posterior but as squared–error loss $\sum_i (y_i - a(x_i, w))^2$ for actions (or approximations) $a(x) = a(x, w)$. According to Section 2.2.2 minimization of error functional (4.104) for data $\{(x_i, y_i) | 1 \leq i \leq n\}$, sampled under the true density $p(x, y|f)$, yields therefore an empirical estimate for the regression function $\int dy\, y\, p(y|x, f)$.

We can choose neural nets as likelihood parameterizations for density estimation problems with prior term (and normalization constraint) explicitly included in the functional $E(\phi)$. The stationarity equation for functional (4.1) then becomes

$$0 = \Phi'_w \mathbf{P}' \mathbf{P}^{-1} N - \Phi'_w \mathbf{K} \phi - \Phi'_w \mathbf{P}' \Lambda_X, \tag{4.105}$$

with matrix of derivatives

$$\begin{aligned} \Phi'_w(k, l, i; x, y) &= \frac{\partial \phi(x, y, w)}{\partial w_{kl,i}} \qquad (4.106) \\ &= \sigma'(v_n) \sum_{l_{n-1}} w_{l_{n-1},n}\, \sigma'(v_{l_{n-1},n-1}) \sum_{l_{n-2}} w_{l_{n-1} l_{n-2},n-1} \\ &\quad \cdots \sum_{l_{i+1}} w_{l_{i+2} l_{i+1},i+2}\, \sigma'(v_{l_{i+1},i+1}) w_{l_{i+1} k,i+1} \sigma'(v_{l_i,i}) z_{l,i-1}, \end{aligned}$$

Ansatz	Functional form	optimized
linear ansatz	$\phi(z) = \sum_l \xi_l B_l(z)$	ξ_l
linear model with interaction	$\phi(z) = \xi_0 + \sum_l \xi_l z_l + \sum_{mn} \xi_{mn} z_m z_n + \cdots$	ξ_0, ξ_l $\xi_{mn}, \cdots$
mixture model	$\phi(z) = \sum \xi_{0,l} B_l(\xi_l, z)$	$\xi_{0,l}$, ξ_l
additive model with interaction	$\phi(z) = \sum_l \phi_l(z_l) + \sum_{mn} \phi_{mn}(z_m z_n) + \cdots$	$\phi_l(z_l)$, $\phi_{mn}(z_m z_n)$, $\cdots$
product ansatz	$\phi(z) = \prod_l \phi_l(z_l)$	$\phi_l(z_l)$
decision trees	$\phi(z) = \sum_l \xi_l \prod_k \Theta(z_{\xi_{lk}} - \xi_{0,lk})$	ξ_l, $\xi_{0,lk}$, ξ_{lk}
projection pursuit	$\phi(z) = \xi_0 + \sum_l \phi_l(\xi_{0,l} + \sum_l \xi_l z_l)$	ϕ_l, ξ_0, $\xi_{0,l}$, ξ_l
neural net (2 layers)	$\phi(z) = \sigma(\sum_l \xi_l \sigma(\sum_k \xi_{lk} z_k))$	ξ_l, ξ_{lk}

Table 4.2: Some possible likelihood parameterizations.

and $\sigma'(v) = d\sigma(v)/dv$. While $\phi(x, y, w)$ is calculated by forward propagating $z = (x, y)$ through the net, defined by weight vector w according to Eq. (4.103), the derivatives Φ' can efficiently be calculated by back–propagation according to Eq. (4.106). Notice that even for diagonal $\mathbf{P}'$ the derivatives are not only needed at data points, but prior and normalization term require derivatives at all x, y. Thus, in practice, terms like $\Phi'\mathbf{K}\phi$ can only be calculated in a relatively poor discretization. (Such prior terms can, for example, be used to create virtual training examples or 'hints' [2,4,161,313,369,404]) Notice, however, that regularization is here not only due to the prior term but follows also from the restrictions implicit in a chosen neural network architecture. In many practical cases a relatively poor discretization of the prior term may thus be sufficient.

Table 4.2 summarizes the discussed approaches.

Chapter 5

Parameterizing priors: Hyperparameters

5.1 Quenched and annealed prior normalization

The quality of nonparametric Bayesian approaches depends mainly on the adequate implementation of problem specific *a priori* information. Especially complex tasks with relatively few training data available, for example, in speech recognition or image reconstruction, require task specific priors. Choosing a simple Gaussian smoothness prior is not sufficient for such tasks. Hence successful practical applications of nonparametric Bayesian methods require flexible methods to adapt priors to the situation under study. Several such techniques will be discussed in this section.

For instance, in the numerical study of Section 6.6, an image completion problem will be studied, where the image of a face is reconstructed from noisy data sampled from that image. In that example hyperparameters represent translational, rotational, and scaling degrees of freedom which determine the specific appearance of a prototypical face used as reference for the reconstruction of the unknown face from noisy image data. In that example the training data contain enough information to determine such hyperparameters. In other situations, (part of the) hyperparameters can be adapted by including a larger set of similar learning tasks. For instance, if several images from the same person have to be reconstructed, hyperparameters like eye color or the ratio of the eye–eye distance to an eye–mouth distance can be estimated using several training images from the same person (cf., for example, [244,475]). Hyperparameters like the specific lighting conditions, scaling, translation, 2D or 3D rotations, or more general face deformations related for instance to specific emotional expressions, e.g., parameterized by an optical flow field [44,233,524],

can then be optimized, using only the training data obtained from a specific view of the face to be reconstructed. This kind of hierarchical determination of hyperparameters is a natural possibility to implement transfer between related learning problems in a nonparametric framework.[1]

Technically speaking, after studying parameters of the likelihood in Section 4, we consider now parameters of the prior density $p(\phi|D_0)$. For Gaussian prior densities this allows, for instance, the parameterization of the mean function and the covariance operator. The parameters of the prior functional, which we will denote by θ, are in a Bayesian context known as *hyperparameters*. Like parameters ξ or fields ϕ, hyperparameters θ are not directly observable and thus part of the hidden variables of a model. Therefore, analogous to priors on ϕ, *hyperpriors* $p(\theta)$ on θ have to be specified. For the corresponding (hyperprior) energies we write $E(\theta)$. The introduction of (one or several levels of) hyperparameters is known as *hierarchical Bayes approach* [174,175].

In the presence of hyperparameters, a prior for ϕ can be decomposed as follows (skipping D_0)

$$p(\phi) = \int d\theta \, p(\phi|\theta)p(\theta). \tag{5.1}$$

While the likelihood does not depend on θ, it follows from Eq. (5.1) that posterior as well as predictive density include an integral over θ. If the integral over θ is low dimensional and can be solved analytically, then one can directly work with the resulting 'marginal' prior $p(\phi)$ and skip the hyperparameters again [68,321,323,474,526,530]. For Gaussian $p(\phi|\theta)$ the resulting $p(\phi)$ is, however, not necessarily Gaussian. Alternatively, one can either use numerical methods like Monte Carlo integration [521,522] or refer to a saddle point approximation to replace integration by maximization of the posterior. When calculating integrals over ϕ and θ in saddle point approximation then, besides the conventional MAP stationarity equations for ϕ, additional MAP stationarity equations result from the derivatives with respect to θ. A joint MAP for ϕ and θ and related approximations will be discussed in Section 5.2.

Sometimes it is convenient to introduce besides the field ϕ a few additional parameters to determine the likelihood. In regression such a parameter can be an x–independent variance σ^2 in the likelihood (3.164). A similar example is the inverse temperature β for mixture models as introduced in Section 6.4 which like ϕ appears in the likelihood and the prior. Such parameters may formally be added to the 'direct' hidden variables ϕ yielding an enlarged $\tilde{\phi}$ and can often be treated analogously to hyperparameters. Clearly, the functional derivatives of the posterior with respect to such 'additional likelihood parameters' then contain terms arising from the derivative of the likelihood.

Within the Frequentist interpretation of error minimization as empirical risk minimization, discussed in Section 2.2.5, hyperparameters θ can be de-

[1] A similar mechanism of transfer between related learning tasks may also be at work in the cognitive system of animals or humans, cf., for example, [137].

termined by minimizing the *empirical generalization error* on a new set of test or validation data D_T, independent of the training data D. Here the empirical generalization error is the pure data term $E_D(\theta) = E_D(\phi^*(\theta))$ of the error functional at ϕ^*, being the optimal ϕ for the full regularized $E(\phi|\theta)$ at θ and for given training data D. Elaborate techniques to determine hyperparameters include cross–validation and bootstrap methods, see also Section 5.4.5.

Within the Bayesian interpretation of error minimization as posterior maximization the introduction of hyperparameters leads to a specific difficulty. The problem arises from the fact that from a Bayesian point of view the error term $E(\theta)$ is interpreted as prior energy for θ, meaning that

$$p(\theta) = \frac{e^{-E(\theta)}}{Z_\theta} \tag{5.2}$$

represents the prior density for θ, with normalization

$$Z_\theta = \int d\theta\, e^{-E(\theta)}. \tag{5.3}$$

Because the joint prior factor for ϕ and θ is given by the product

$$p(\phi, \theta) = p(\phi|\theta)p(\theta), \tag{5.4}$$

one finds

$$p(\phi|\theta) = \frac{e^{-E(\phi|\theta)}}{Z_\phi(\theta)}. \tag{5.5}$$

Hence the ϕ–dependent part of the energy represents a *conditional (or quenched) prior energy*, denoted by $E(\phi|\theta)$. As the conditional normalization

$$Z_\phi(\theta) = \int d\phi\, e^{-E(\phi|\theta)} \tag{5.6}$$

is in general θ–dependent, a normalization term

$$E_N(\theta) = \ln Z_\phi(\theta) \tag{5.7}$$

must also be included in the error functional when minimizing with respect to θ. Notice that in general $E(\phi|\theta)$ depends on θ, even if $p(\theta)$ is uniform and does not appear explicitly in the error functional.

It is interesting to see what happens if $p(\phi, \theta)$ of Eq. (5.4) is expressed in terms of *joint (or annealed) prior energy* $E(\phi, \theta)$ as follows

$$p(\phi, \theta) = \frac{e^{-E(\phi,\theta)}}{Z_{\phi,\theta}}. \tag{5.8}$$

Then the joint normalization

$$Z_{\phi,\theta} = \int d\phi\, d\theta\, e^{-E(\phi,\theta)} \tag{5.9}$$

is independent of ϕ and θ and can be skipped from the functional. However, in that case the term $E(\phi,\theta)$ can not easily be related to the prior $p(\theta) = \int d\phi p(\phi,\theta)$, and in practice it is typically much easier to build a model separately for $p(\theta)$ and $p(\phi|\theta)$, than for the joint density $p(\phi,\theta)$.

The two ways of expressing $p(\phi,\theta)$ by a joint or conditional energy, respectively, correspond to the distinction of systems with annealed or quenched disorder like spin glasses (see [128,294,338] and Appendix B.5). The two formulations are equivalent if the joint density factorizes. However, in that case θ and ϕ are independent, and θ does not parameterize the density of ϕ.

Numerically the need to calculate $Z_\phi(\theta)$ can be disastrous because normalization factors $Z_\phi(\theta)$ are often extremely high dimensional (functional) integrals which can be very difficult to calculate, in contrast to the normalization of $p(y|x,\phi)$ as integral over y.

There are, however, situations for which $Z_\phi(\theta)$ remains θ–independent. Let $p(\phi,\theta)$ stand for a Gaussian specific prior $p(\phi,\theta|\tilde{D}_0)$ (with the normalization condition factored out as in Eq. (2.101)). Then, because the normalization of a Gaussian is independent of its mean, parameterizing the mean $t = t(\theta)$ results in a θ–independent $Z_\phi(\theta)$.

Besides their mean, Gaussian processes are characterized by their covariance operators $\mathbf{K}^{-1}$. Because the normalization only depends on $\det \mathbf{K}$, a second possibility yielding θ–independent $Z_\phi(\theta)$ are parameterized transformations of the form $\mathbf{K} \to \mathbf{OKO}^{-1}$ with orthogonal $\mathbf{O} = \mathbf{O}(\theta)$. Indeed, such transformations do not change the determinant $\det \mathbf{K}$. They are only non–trivial in more than one dimension. (See Section 5.4.4.)

For general parameterizations of density estimation problems, however, the normalization term $E_N(\theta) = \ln Z_\phi(\theta)$ of Eq. (5.7) must be included. The only way to get rid of that normalization term is to introduce ad hoc a *compensating hyperprior*

$$p(\theta) \propto Z_\phi(\theta), \tag{5.10}$$

resulting in an error term $E(\theta) = -\ln Z_\phi(\theta)$ compensating the normalization term $E_N(\theta)$.

In the general case we have to consider the functional

$$E(\phi,\theta) = -\langle \ln P(\phi)|\, N\rangle + \langle P(\phi)|\, \Lambda_X\rangle + E(\phi|\theta) + E(\theta) + \ln Z_\phi(\theta). \tag{5.11}$$

The stationarity conditions have the form

$$\frac{\delta E(\phi|\theta)}{\delta \phi} = \mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N - \mathbf{P}'(\phi)\Lambda_X, \tag{5.12}$$

$$\frac{\partial E(\phi|\theta)}{\partial \theta} = -\mathbf{Z}' Z_\phi^{-1}(\theta) - E'(\theta), \tag{5.13}$$

with

$$\mathbf{Z}'(l,k) = \delta(l-k)\frac{\partial Z_\phi(\theta)}{\partial \theta_l}, \quad E'(\theta,l) = \frac{\partial E(\theta)}{\partial \theta_l}. \tag{5.14}$$

For a compensating hyperprior $E(\theta) = -\ln Z_\phi(\theta)$ the right hand side of Eq. (5.13) vanishes by definition.

Finally, we want to remark that in case evaluation of the function $p(\phi,\theta)$ is much cheaper than calculating the gradient (5.13), minimization methods not using the gradient should be considered, like, for example, the downhill simplex method [410]. Similarly, parallel optimization methods, like genetic algorithms [99, 167, 218, 274, 339, 346, 444], and homotopy methods, like simulated annealing [1, 98, 149, 261, 338, 415, 428, 532–534], can be used with or without gradient information.

5.2 Saddle point approximations and hyperparameters

5.2.1 Joint MAP

To include hyperparameters θ, the straightforward extension of a MAP approach for fields ϕ (or parameters ξ) is a MAP with respect to all hidden variables ϕ and θ [347]. Indeed, to calculate the predictive density (where one can write $p(\theta|D_0)$ for $p(\theta)$)

$$p(y|x,D,D_0) = \frac{1}{p(y_D|x_D,D_0)} \int d\phi\, p(y|x,\phi) \int d\theta \underbrace{p(y_D|x_D,\phi)p(\phi|D_0,\theta)p(\theta)}_{\propto p(\phi,\theta|D,D_0),\ \max \text{ w.r.t. } \theta \text{ and } \phi} \tag{5.15}$$

in saddle point approximation, the joint posterior $p(\phi,\theta|D,D_0)$ must be maximized with respect to ϕ and θ, assuming a slowly varying $p(y|x,\phi)$ at the stationary point.

This simultaneous maximization with respect to both variables is consistent with the usual asymptotic justification of a saddle point approximation. For example, for a function $f(\phi,\theta)$ of two (one–dimensional) variables ϕ, θ

$$\int d\phi\, d\theta\, e^{-\beta f(\phi,\theta)} \approx e^{-\beta f(\phi^*,\theta^*) - \frac{1}{2}\ln\det(\frac{\beta \mathbf{H}}{2\pi})} \tag{5.16}$$

for large enough β (and a unique maximum). Here $f(\phi^*,\theta^*)$ denotes the joint minimum and $\mathbf{H}$ the Hessian of f with respect to ϕ and θ. For a Gaussian $p(\phi|\theta)$ with θ–dependent determinant of the covariance, the function has the form $f(\phi,\theta) = E(\phi,\theta) + (1/(2\beta))\ln\det(\beta\mathbf{K}(\theta)/(2\pi))$, where both terms are relevant for the minimization of f with respect to θ. For large β, however, the second term becomes small compared to the first one. (Of course, there is the possibility that a saddle point approximation is not adequate for the

θ integration. For example, hard constraints, like the condition of a positive definite covariance, restrict the θ–values to certain regions while a Gaussian decreases with infinite tails. Similarly, the MAP for ϕ may become a poor approximation for small likelihoods $p(y|x,\phi)$ near the border where the positivity constraint restricts the region of allowed likelihoods. Hence for density estimation problems one would expect a MAP to show a tendency towards overfitting, choosing likelihoods too small in regions where no data are available.)

5.2.2 Stepwise MAP

As alternative to a joint saddle point approximation, one can perform the two integrals stepwise. This seems especially useful if one integral can be calculated analytically. Consider, for example,

$$\int d\phi\, d\theta\, e^{-\beta f(\phi,\theta)} \approx \int d\theta\, e^{-\beta f(\phi^*(\theta),\theta) - \frac{1}{2}\ln\det(\frac{\beta}{2\pi}\frac{\delta^2 f(\phi^*(\theta))}{\delta\phi^2})}, \qquad (5.17)$$

which is exact for a Gaussian ϕ–integral. One sees now that minimizing the complete negative exponent $\beta f(\phi^*,\theta) + \frac{1}{2}\ln\det(\beta(\delta^2 f/\delta\phi^2)/2\pi)$ with respect to θ is different from minimizing $f(\phi,\theta)$ in (5.16), if the second derivative of f with respect to ϕ depends on θ (which is not the case for a Gaussian θ integral). Again this additional term becomes asymptotically negligible, i.e., for β large enough. Stepwise approaches like (5.17) can be used, for example, if one (Gaussian) integration can be performed analytically and can lead to somewhat simpler stationarity equations for θ–dependent covariances [521]. There are also cases where hyperparameters can be integrated out exactly typically leaving a non–Gaussian integral in ϕ which then has to be solved by MAP or Monte Carlo integration [68,474,526,530]. (See also Section 5.4.5.)

5.2.3 Pointwise approximation

Let us consider the case of Gaussian regression in some more detail. (The following discussion also applies to density estimation if, as in (5.17), the Gaussian first step integration is replaced by a saddle point approximation including the normalization factor. This requires the calculation of the determinant of the Hessian.) Consider the two step procedure for Gaussian regression

$$p(y|x,D,D_0) = \frac{1}{p(y_D|x_D,D_0)}\int d\theta\; \underbrace{p(\theta)\underbrace{\int d\phi\, p(y|x,\phi)p(y_D|x_D,\phi)p(\phi|D_0,\theta)}_{\text{exact}}}_{p(\theta)p(y,y_D|x,x_D,D_0,\theta)\propto p(y,\theta|x,D,D_0)\ \max\text{ w.r.t. }\theta},$$

$$= \underbrace{\int d\theta \underbrace{p(\theta|D,D_0)}_{\propto \text{exact}} \underbrace{p(y|x,D,D_0,\theta)}_{\text{exact}}}_{p(y,\theta|x,D,D_0),\ \max \text{w.r.t.}\ \theta} \tag{5.18}$$

where in a first step $p(y, y_D|x, x_D, D_0, \theta)$ can be calculated analytically and in a second step the θ integral is performed by Gaussian approximation around a stationary point. Instead of maximizing the joint posterior $p(\phi, \theta|D, D_0)$ in Eq. (5.15) with respect to ϕ and θ this approach performs the ϕ–integration analytically and maximizes $p(y, \theta|x, D, D_0)$ with respect to θ. The disadvantage of this approach is that the maximization must be performed for each pair y, x separately. This is not the case for the joint or stepwise MAP discussed in the previous Sections 5.2.1 and 5.2.2.

5.2.4 Marginal posterior and empirical Bayes

Assuming a slowly varying $p(y|x, D, D_0, \theta)$ at the stationary point one can try a saddle point approximation of the predictive density (5.18) by maximizing the ϕ–marginalized posterior

$$p(\theta|D, D_0) = \int d\phi\, p(\phi, \theta|D, D_0). \tag{5.19}$$

In Gaussian regression where the ϕ–integration in the predictive density (5.18) can be performed exactly,

$$p(y|x, D, D_0) = \int d\theta \underbrace{\underbrace{p(\theta|D, D_0)}_{\substack{\text{exact} \\ \max \text{w.r.t.}\ \theta}} \underbrace{p(y|x, D, D_0, \theta)}_{\text{exact}}}, \tag{5.20}$$

the analytical expression for $p(y|x, D, D_0, \theta)$ for some fixed θ is given by Eq. (3.265) of Section 3.8.2. Analogously, the posterior density $p(\theta|D, D_0)$ can be obtained from the likelihood of θ and a specified hyperprior $p(\theta)$

$$p(\theta|D, D_0) = \frac{p(y_D|x_D, D_0, \theta)p(\theta)}{p(y_D|x_D, D_0)}. \tag{5.21}$$

Thus in case the θ–likelihood can be calculated analytically, the θ–integral in (5.20) can be calculated in saddle point approximation by maximizing the posterior for θ with respect to θ. For a uniform $p(\theta)$ the optimal θ^* is obtained by maximizing the θ–likelihood. The ϕ–marginalized θ–likelihood $p(y_D|x_D, D_0, \theta)$ is also known as *marginalized likelihood*, or as evidence of the data y_D conditional on x_D and θ. For Gaussian regression also the θ-likelihood in Eq. (5.21) can be integrated analytically. Analogous to the integration of

Eq. (3.254) in Section 3.8.2, yielding Eq. (3.265), one finds [507,521,522]

$$\begin{aligned} p(y_D|x_D, D_0, \theta) &= \int d\phi\, p(y_D|x_D, \phi)\, p(\phi|D_0, \theta) \\ &= \int d\phi\, e^{-\frac{1}{2}\sum_{i=0}^{n}\langle \phi - t_i | \mathbf{K}_i | \phi - t_i\rangle + \frac{1}{2}\sum_{i=0}^{n} \ln \det_i(\mathbf{K}_i/2\pi)} \\ &= e^{-\widetilde{E} + \frac{1}{2}\ln \det_D \widetilde{\mathbf{K}} - \frac{n}{2}\ln(2\pi)}, \end{aligned} \tag{5.22}$$

where $\widetilde{E} = \frac{1}{2}\left\langle t_D - t_0 \middle| \widetilde{\mathbf{K}} \middle| t_D - t_0 \right\rangle$ and, analogous to Eq. (3.274), $\widetilde{\mathbf{K}} = (\mathbf{K}_D^{-1} + \mathbf{K}_{0,DD}^{-1}(\theta))^{-1} = \mathbf{K}_D - \mathbf{K}_D\mathbf{K}^{-1}\mathbf{K}_D$, with $\mathbf{K} = \sum_{i=0}^{n}\mathbf{K}_i$, and $\det_D$ the determinant in data space. In cases where the marginalization over ϕ, necessary to obtain the evidence, can not be performed analytically and all ϕ–integrals are calculated in saddle point approximation, we get the same result as for a direct simultaneous MAP for ϕ and θ for the predictive density as indicated in (5.15).

The selection of hyperparameter values by maximizing the ϕ–marginalized θ–likelihood, or the ϕ–marginalized posterior of θ, instead of performing the hyperparameter integral, is closely related to the *ML–II prior approach* [33, 34,175], the *generalized maximum likelihood approach* [505], and the *evidence framework* [52,185,186,317–321,323,459,466,468]. Similarly, also the *empirical Bayes* approach selects hyperparameters by maximizing the marginalized likelihood [33, 34, 73, 354, 420–424]. In contrast to our setting, an empirical Bayes approach assumes a compound sampling model. In a compound sampling model the hidden variables ϕ or parameters ξ, respectively, are drawn for each trial (or a group of trials) separately, while we assume the state of Nature (ϕ, θ) to be fixed for all training and test data. Thus in our situation the pure states are $p(y|x, \phi)$ while in a compound sampling model the pure states are $p(y|x, \theta) = \int d\phi\, p(y|x, \phi) p(\phi|\theta)$. In particular, the evidence of n independent data, obtained from the same unknown state of Nature, is

$$p(y_D|x_D, D_0) = \int d\theta\, p(\theta) \underbrace{\int d\phi \left(\prod_i p(y_i|x_i, \phi)\right) p(\phi|D_0, \theta)}_{p(y_D|x_D, D_0, \theta)}, \tag{5.23}$$

while in a compound sampling model for each new data point (x_i, y_i) a new ϕ is drawn from $p(\phi|D_0, \theta)$ with the same unknown θ

$$p(y_D|x_D, D_0) = \int d\theta\, p(\theta) \underbrace{\left(\prod_i \int d\phi_i\, p(y_i|x_i, \phi_i) p(\phi_i|D_0, \theta)\right)}_{p(y_D|x_D, D_0, \theta)}. \tag{5.24}$$

In both cases a saddle point approximation for θ is obtained by maximizing the marginalized likelihood (or the marginalized posterior) which is, however, different for the two models.

As the roles of ϕ or ξ, respectively, are different in a fixed state model (5.23) and in a compound sampling model (5.24), it is useful to switch notation for the compound sampling model. We replace ϕ in (5.24) by $\tilde{\phi}$ or parameters ξ by $\tilde{\xi}$, respectively, which represent 'nuisance' variables of the likelihood to be integrated out. Furthermore, depending on whether we choose a nonparametric model (nonparametric empirical Bayes) or a parametric model (parametric empirical Bayes) for the probability density of $\tilde{\phi}$ or $\tilde{\xi}$, respectively, we replace θ by ϕ or by ξ, respectively. For instance, for a completely parametric compound sampling model we write $p(y|x,\xi) = \int d\tilde{\xi}\, p(y|x,\tilde{\xi})p(\tilde{\xi}|\xi)$ while a corresponding nonparametric approach would be $p(y|x,\phi) = \int d\tilde{\xi}\, p(y|x,\tilde{\xi})p(\tilde{\xi}|\phi)$. Choosing a completely unrestricted nonparametric model for both factors under the integral does not result in a more general approach, as both $p(y|x,\tilde{\phi})$ and $p(y|x,\phi) = \int d\tilde{\phi}\, p(y|x,\tilde{\phi})p(\tilde{\phi}|\phi)$ are then completely free.

It is interesting to ask how in a fixed state model a single draw of ϕ can be sufficient to find a reasonable MAP solution for θ. (A compound sampling model would provide us with a larger sample of ϕ to guess the hyperparameter θ.) To obtain enough information on θ from a single draw is indeed possible if we consider nonparametric models. The reason is that drawing one function ϕ is equivalent to drawing all the function values $\phi(x,y)$. This collection of function values can contain enough information to find good approximations for a few real numbers θ. For example, one can expect to find a good approximation for the periodicity interval θ of a template function $t(x) = t(x+\theta)$ in a Gaussian prior process, if x varies over many periods and the prior process is not too noisy, i.e., has a small covariance. Indeed, the similarity of a nonparametric fixed state model and a compound sampling model can be seen easily if all $(x_i,\, y_i)$-pairs in the data are different. Then we can write in Eq. (5.23) $p(y_D|x_D,D_0,\theta) = \int(\prod_i d\phi_i)\prod_i (p(y_i|x_i,\phi_i))\; p(\phi_1,\cdots\phi_n|D_0,\theta)$. This is a compound sampling model where for given θ the random numbers $\phi_i = \phi(x_i,y_i)$ are dependent, while in Eq. (5.24) the functions $\phi_i = \phi_i(x,y)$ are independent for given θ, $p(\phi_1,\cdots\phi_n|D_0,\theta) = \prod_i p(\phi_i|D_0,\theta)$.

5.2.5 Some variants of stationarity equations

Finally let us compare the stationarity equations for hyperparameters θ resulting from the different approaches for a Gaussian prior with θ–dependent mean $t_0(\theta)$, inverse covariance $\mathbf{K}_0(\theta)$, and hyperprior $p(\theta)$ in the case of Gaussian regression. Setting the derivative of the joint posterior $p(\phi,\theta|D,D_0)$ with respect to θ to zero and using

$$\frac{\partial}{\partial\theta}\ln\det\mathbf{K} = \frac{\partial}{\partial\theta}\mathrm{Tr}\,\ln\mathbf{K} = \mathrm{Tr}\left(\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta}\right) \tag{5.25}$$

yields

$$0 \;=\; \left\langle \frac{\partial t_0}{\partial\theta}\,\middle|\,\mathbf{K}_0\,\middle|\,t_0-\phi\right\rangle + \frac{1}{2}\left\langle \phi-t_0\,\middle|\,\frac{\partial\mathbf{K}_0(\theta)}{\partial\theta}\,\middle|\,\phi-t_0\right\rangle$$

$$-\frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_0^{-1}\frac{\partial \mathbf{K}_0}{\partial\theta}\right) - \frac{1}{p(\theta)}\frac{\partial p(\theta)}{\partial\theta}. \tag{5.26}$$

This equation has to be solved simultaneously with the stationarity equation for ϕ. While this form of a MAP approach is in principle easy to adapt to general density estimation problems, for θ–dependent covariance determinants it can sometimes be difficult to calculate the derivative of the determinant of $\mathbf{K}_0$. On the other hand, maximizing the ϕ–marginalized posterior $p(\theta|D, D_0)$ of Eq. (5.19) as discussed in Section 5.2.4 requires only the calculation of the derivative of the determinant of the $(\tilde{n} \times \tilde{n})$–matrix $\widetilde{\mathbf{K}}$ of Eq. (3.220)

$$\begin{aligned} 0 &= \left\langle \frac{\partial t_0}{\partial\theta}\middle|\widetilde{\mathbf{K}}\middle|t_0 - t_D\right\rangle + \frac{1}{2}\left\langle t_D - t_0\middle|\frac{\partial\widetilde{\mathbf{K}}}{\partial\theta}\middle|t_D - t_0\right\rangle \\ &\quad -\frac{1}{2}\mathrm{Tr}\left(\widetilde{\mathbf{K}}^{-1}\frac{\partial\widetilde{\mathbf{K}}}{\partial\theta}\right) - \frac{1}{p(\theta)}\frac{\partial p(\theta)}{\partial\theta}. \end{aligned} \tag{5.27}$$

The solution of the corresponding 'ϕ'–equation (3.207), resulting from the functional derivative of the posterior with respect to ϕ, is given in Eq. (3.219) as $\phi^* = t_0 + \mathbf{K}_0^{-1}\widetilde{\mathbf{K}}(t_D - t_0)$. Evaluated at this ϕ^*, the first term of the 'θ'–Eq. (5.26), which does not contain derivatives of the inverse covariances, becomes equal to the first term of the alternative 'θ'–Eq. (5.27). The last terms of Eqs. (5.26) and (5.27) are always identical. The derivative of $\widetilde{\mathbf{K}}$ with respect to θ in the second term is calculated in data space, while the corresponding derivative of $\mathbf{K}_0$ in Eq. (5.26) lives in the full ϕ–space. Typically, the data–independent $\mathbf{K}_0$, defined on the full space, has a more regular structure than the data–dependent $\widetilde{\mathbf{K}}$, defined on data space. Thus at least for one or two dimensional x, a straightforward numerical solution of Eq. (5.26) by discretizing x can also be a good choice for Gaussian regression problems.

Analogously, from Eq. (3.265) follows for maximizing $p(y, \theta|x, D, D_0)$ with respect to θ

$$\begin{aligned} 0 &= \left\langle \frac{\partial t}{\partial\theta}\middle|\mathbf{K}_y\middle|t - y\right\rangle + \frac{1}{2}\left\langle y - t\middle|\frac{\partial\mathbf{K}_y}{\partial\theta}\middle|y - t\right\rangle \\ &\quad -\frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_y^{-1}\frac{\partial\mathbf{K}_y}{\partial\theta}\right) - \frac{1}{p(\theta|D, D_0)}\frac{\partial p(\theta|D, D_0)}{\partial\theta}, \end{aligned} \tag{5.28}$$

which is y– and x–dependent. Such an approach can be considered if only specific test data x, y are of interest.

We may remark that also in Gaussian regression the θ–integral may be quite different from a Gaussian integral, so a saddle point approximation does not necessarily give satisfactory results [500, 501]. Due to the appearance of Jacobian determinants, a saddle point approximation is not invariant under transformations of the integration variables. In cases one encounters problems

one can thus look for transformations $\int f(\theta)d\theta = \int \det(\partial\theta/\partial\theta')f(\theta(\theta'))d\theta'$ for which the integrand is better approximated by a Gaussian. The variability of saddle point approximations results from the freedom to add terms which vanish asymptotically but remain finite in the nonasymptotic region. Similar effects are known in quantum many body theory, where, for instance, a Hartree approximation, a Hartree–Fock approximation, and a Hartree–Fock–Bogoliubov approximation with pairing fields can all be obtained from the same type of stationary phase approximation using just different decompositions of the Hamiltonian (see, for example, Chapter 7 in [365], in particular Problem 7.1). Thus finding the right parameterization for a saddle point approximation is still an art.

5.3 Adapting prior means

5.3.1 General considerations

A prior mean or template function t represents a prototype (reference function or base line) for ϕ. It may be a typical, expected pattern in time series prediction or a reference image in image reconstruction. Consider, for example, the task of completing an image ϕ given some pixel values (training data) [297]. Expecting the image to be that of a face, the template function t can be chosen to be some prototypical image of a face. We have seen in Section 3.7 that a single template t can be eliminated for Gaussian (specific) priors by solving for $\phi - t$ instead for ϕ. Restricting to only a single template may be a very bad choice. Indeed, faces, for example, appear on images in many variations, like in different scales, translated, rotated, under various illuminations and typical deformations. We can describe such variations by a family of templates $t(\theta)$, the parameter θ describing scaling, translations, rotations, and more general deformations. Thus we expect a function to be similar to only one of the templates $t(\theta)$ and use hyperparameters θ to implement a (soft, probabilistic) OR, approximating $t(\theta_1)$ OR $t(\theta_2)$ OR $\cdots$ (See also [291–295]).

A (soft, probabilistic) AND of approximation conditions, on the other hand, is implemented by adding error terms. For example, classical error functionals where data and prior terms are added correspond to an approximation of training data AND *a priori* data.

Similar considerations apply for *model selection*. We could, for example, expect ϕ to be well approximated by a neural network or a decision tree, respectively. In that case $t(\theta)$ spans a space of neural networks or decision trees. Finally, let us emphasize again that the great advantage and practical feasibility of adaptive templates for regression problems comes from the fact that no additional normalization terms have to be added to the error functional.

5.3.2 Density estimation and nonparametric boosting

Density estimation with adaptive means for Gaussian prior factors and hyperparameter energy $E(\theta)$ yields an error functional

$$E(\phi,\theta) = -\langle \ln P(\phi)|\, N\rangle + \frac{1}{2}\langle \phi - t(\theta)\,|\mathbf{K}|\,\phi - t(\theta)\rangle + \langle P(\phi)|\,\Lambda_X\rangle + E(\theta). \tag{5.29}$$

Defining

$$\mathbf{t}'(l;x,y) = \frac{\partial t(x,y;\theta)}{\partial \theta_l}, \tag{5.30}$$

the stationarity equations of (5.29), obtained from the functional derivatives with respect to ϕ and hyperparameters θ, become

$$\mathbf{K}(\phi - t) = \mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N - \mathbf{P}'(\phi)\Lambda_X, \tag{5.31}$$

$$\mathbf{t}'\mathbf{K}(\phi - t) = E'(\theta). \tag{5.32}$$

Inserting Eq. (5.31) in Eq. (5.32) gives

$$\mathbf{t}'\mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N = \mathbf{t}'\mathbf{P}'(\phi)\Lambda_X + E'(\theta). \tag{5.33}$$

For $\phi = t(\theta)$, assuming t fulfills the non–negativity and normalization conditions, Eq. (5.33) has the form of the parametric stationarity equation (4.4). A non–vanishing prior term in (4.4) can be identified with the term $E'(\theta)$. This shows that parametric methods can be considered as 'deterministic' limits of (prior mean) hyperparameter approaches where the Gaussian prior on ϕ approaches a δ–function setting $\phi = t(\theta)$. *In particular, a parametric solution can thus serve as reference template t, to be used within a specific prior factor.* In this way, a nonparametric approach can be used to extend a parametric model. This may be called 'nonparametric boosting of a parametric model'. Similarly, such a parametric solution is a natural initial guess for a nonparametric ϕ when solving a stationarity equation by iteration.

If working with parameterized $\phi(\xi)$ extra prior terms, Gaussian in some function $\psi(\xi)$, can be included as discussed in Section 4.2. Then, analogous to templates t for ϕ, also parameter templates t_ψ can be adapted with hyperparameters θ_ψ. Furthermore, prior terms $E(\theta)$ and $E(\theta_\psi)$ for the hyperparameters θ, θ_ψ can be added. Including such additional error terms yields

$$\begin{aligned} E(\theta,\theta_\psi,\phi(\xi),\psi(\xi)) = & -\langle \ln P(\,\phi(\xi)\,)|\, N\rangle + \langle P(\,\phi(\xi)\,)|\,\Lambda_X\rangle \\ & + \frac{1}{2}\langle \phi(\xi) - t(\theta)\,|\mathbf{K}|\,\phi(\xi) - t(\theta)\rangle \\ & + \frac{1}{2}\langle \psi(\xi) - t_\psi(\theta_\psi)\,|\mathbf{K}_\psi|\,\psi(\xi) - t_\psi(\theta_\psi)\rangle \\ & + E(\theta) + E(\theta_\psi), \end{aligned} \tag{5.34}$$

and Eqs. (5.31) and (5.32) change to

$$\begin{aligned} \mathbf{\Phi}'\mathbf{K}(\phi - t) + \mathbf{\Psi}'\mathbf{K}_\psi(\psi - t_\psi) &= \mathbf{P}'_\xi \mathbf{P}^{-1} N - \mathbf{P}'_\xi \Lambda_X, & (5.35) \\ \mathbf{t}'\mathbf{K}(\phi - t) &= E'(\theta), & (5.36) \\ \mathbf{t}'_\psi \mathbf{K}_\psi(\psi - t_\psi) &= E'(\theta_\psi), & (5.37) \end{aligned}$$

where $\mathbf{t}'_\psi$, $E'(\theta_\psi)$, $E'(\theta)$ denote derivatives with respect to the parameters θ_ψ or θ, respectively. Parameterizing $E(\theta)$ and $E(\theta_\psi)$, the process of introducing hyperparameters can be iterated.

5.3.3 Unrestricted variation

To get some understanding of the approach (5.29) let us consider the extreme example of completely unrestricted t–variations. In that case the template function $t(x, y)$ itself represents the hyperparameter and we choose $E'(\theta) = 0$. (Function hyperparameters or hyperfields like such a $\theta(x, y) = t(x, y)$ are discussed in detail in Section 5.6.) Then, $\mathbf{t}' = \mathbf{I}$ and Eq. (5.32) gives $\mathbf{K}(\phi - t) = 0$ (which for invertible $\mathbf{K}$ is solved uniquely by $t = \phi$), resulting according to Eq. (3.159) in

$$\Lambda_X = N_X. \tag{5.38}$$

The case of a completely free prior mean t is therefore equivalent to a situation without prior. Indeed, as $\mathbf{t}' = \mathbf{I}$ and $\mathbf{P}$ is diagonal, projection of Eq. (5.33) into the (x–)data space by $\mathbf{I}_D$ of Eq. (3.182) yields for invertible $\mathbf{P}'$

$$P_D = \mathbf{\Lambda}^{-1}_{X,D} N, \tag{5.39}$$

where $\mathbf{\Lambda}_{X,D} = \mathbf{I}_D \mathbf{\Lambda}_X \mathbf{I}_D$ is invertible in data space and $P_D = \mathbf{I}_D P$. Thus for x_i for which y_i are available

$$P(x_i, y_i) = \frac{N(x_i, y_i)}{N_X(x_i)} \tag{5.40}$$

is concentrated on the data points. Comparing this with solutions of Eq. (3.158), for $\theta = t$ fixed, we see that adaptive means tend to lower the influence of prior terms.

5.3.4 Regression

Consider now the case of regression according to functional (3.186) with an adaptive template $t_0(\theta)$. The system of stationarity equations, for the regression function $\phi(x)$ and hyperparameter θ, becomes for a uniform hyperprior $E(\theta)$

$$\begin{aligned} \mathbf{K}_0(\phi - t_0) &= \mathbf{K}_D(t_D - \phi), & (5.41) \\ \mathbf{t}'_0 \mathbf{K}_0(\phi - t_0) &= 0. & (5.42) \end{aligned}$$

It will be useful to insert Eq. (5.41) in Eq. (5.42), yielding

$$0 = \mathbf{t}_0' \mathbf{K}_D (\phi - t_D). \tag{5.43}$$

For fixed θ, and thus given t_0, Eq. (5.41) is solved by the template average t of Eq. (3.193)

$$\phi = (\mathbf{K}_0 + \mathbf{K}_D)^{-1} (\mathbf{K}_0 t_0 + \mathbf{K}_D t_D) = t, \tag{5.44}$$

so that Eq. (5.42) or Eq. (5.43), respectively, become

$$0 = \mathbf{t}_0' \mathbf{K}_0 (t - t_0), \tag{5.45}$$

$$0 = \mathbf{t}_0' \mathbf{K}_D (t - t_D). \tag{5.46}$$

It is now interesting to note that if we replace in Eq. (5.46) the full template average t by t_0 we get

$$0 = \mathbf{t}_0' \mathbf{K}_D (t_0 - t_D). \tag{5.47}$$

Replacing $t(\theta)$ by $\phi(\xi)$, this is equivalent to the stationarity equation

$$0 = \Phi' \mathbf{K}_D (\phi - t_D), \tag{5.48}$$

with derivative matrix $\Phi' = \partial\phi/\partial\xi$ as in Eq. (4.3), of an error functional

$$E_D(\phi(\xi)) = \frac{1}{2} \langle \phi(\xi) - t_D \,|\mathbf{K}_D|\, \phi(\xi) - t_D \rangle \tag{5.49}$$

without prior terms but with parameterized $\phi(\xi)$, e.g., a neural network. The approximation $\phi = t = t_0$ can be interpreted as limit $\lambda \to \infty$,

$$\lim_{\lambda \to \infty} \phi = \lim_{\lambda \to \infty} t = t_0, \tag{5.50}$$

after replacing $\mathbf{K}_0$ by $\lambda \mathbf{K}_0$ in Eq. (5.44). Setting $\phi = t_0$ can then be used as initial guess ϕ^0 for an iterative solution for ϕ. For existing $\mathbf{K}_0^{-1}$, the function $\phi = t_0$ is also obtained after one iteration step of the iteration scheme $\phi^{(i)} = t_0 + \mathbf{K}_0^{-1} \mathbf{K}_D (t_D - \phi^{(i-1)})$, when solving Eq. (5.41) starting with initial guess $\phi^{(0)} = t_D$.

For comparison with Eqs. (5.46), (5.47), and (5.48), we give the stationarity equations for parameters ξ for a parameterized regression functional as in (4.2), where compared to Eq. (5.49) an additional prior term with hyperparameters is included,

$$\begin{aligned} E(\phi(\xi), \theta) \;=\; & \frac{1}{2} \langle \phi(\xi) - t_D \,|\mathbf{K}_D|\, \phi(\xi) - t_D \rangle \\ & + \frac{1}{2} \langle \phi(\xi) - t_0(\theta) \,|\mathbf{K}_0(\theta)|\, \phi(\xi) - t_0(\theta) \rangle , \end{aligned} \tag{5.51}$$

which are

$$0 = \Phi' \mathbf{K}_D (\phi - t_D) + \Phi' \mathbf{K}_0 (\phi - t_0). \tag{5.52}$$

Let us now compare the various regression functionals we have met up to now. The non–parametric, regularized regression functional $E(\phi)$ of Eq. (3.186) implements prior information explicitly by a regularization term.

A parametric, regularized functional $E(\phi(\xi))$ of the form (4.2) corresponds to a functional of the form (5.51) for θ fixed. It imposes restrictions on the regression function ϕ in two ways, by choosing a specific parameterization and by including an explicit prior term. If the number of data is large enough, compared to the flexibility of the parameterization, the data term of $E(\phi(\xi))$ alone can have a unique minimum. Then, at least technically, no additional prior term would be required. This corresponds to the classical error minimization methods used typically for parametric approaches.

The regularized functional with hyperparameters, $E(\phi, \theta)$ of Eq. (5.29), implements effectively weaker prior restrictions than a corresponding functional $E(\phi)$ without hyperparameters, as in Eq. (3.157). The prior term corresponds to a *soft restriction* of ϕ to the space of functions spanned by the parameterized templates $t(x, y; \theta)$ (corresponding to $t_0(x, \theta)$ in regression). In the limit where the parameterization of $t(\theta)$, or $t_0(\theta)$, respectively, is rich enough to allow $t(\theta^*) = \phi^*$ at the stationary point the prior term vanishes completely.

The parameterized and regularized functional $E(\phi(\xi), \theta)$ of Eq. (5.51), including prior parameters θ, implements prior information explicitly by a regularization term and implicitly by the parameterization of $\phi(\xi)$. The explicit prior term vanishes if $t_0(\theta^*) = \phi(\xi^*)$ at the stationary point. The functional combines a *hard restriction* of ϕ with respect to the space spanned by the parameterization $\phi(\xi)$ and a soft restriction of ϕ with respect to the space spanned by the parameterized $t_0(\theta)$. Finally, the parameterized and non–regularized functional $E_D(\phi(\xi))$ of Eq. (5.49) implements prior information only implicitly by parameterizing $\phi(\xi)$. In contrast to the functionals $E(\phi, \theta)$ and $E(\phi(\xi), \theta)$ it implements only a hard restriction for ϕ. Table 5.1 summarizes the discussion.

Functional	Eq.	prior implemented
$E(\phi)$	(3.186)	explicitly
$E(\phi(\xi))$	(4.2)	explicitly and implicitly
$E(\phi, \theta)$	(5.29)	explicitly
		no prior for $t(\theta^*) = \phi^*$
$E_D(\phi(\xi))$	(5.49)	implicitly
$E(\phi(\xi), \theta)$	(5.51)	explicitly and implicitly
		no explicit prior for $t_0(\theta^*) = \phi(\xi^*)$

Table 5.1: Some variants of implicit and explicit implementation of *a priori* information.

5.4 Adapting prior covariances

5.4.1 General case

Parameterizing covariances $\mathbf{K}^{-1}$ is often desirable in practice. It includes, for example, adapting the trade–off between data and prior terms, the selection between different symmetries, smoothness measures, or in the multidimensional situation the determination of directions with low variance. As far as the normalization depends on $\mathbf{K}(\theta)$ one has to consider the error functional

$$\begin{aligned} E(\phi,\theta) &= -\langle \ln P(\phi)|\, N\rangle + \frac{1}{2}\langle \phi - t\,|\mathbf{K}(\theta)|\,\phi - t\rangle \\ &\quad + \langle P(\phi)|\,\Lambda_X\rangle + \ln Z_\phi(\theta) + E(\theta), \end{aligned} \tag{5.53}$$

with

$$Z_\phi(\theta) = (2\pi)^{\frac{d}{2}}(\det \mathbf{K}(\theta))^{-\frac{1}{2}}, \tag{5.54}$$

for a d–dimensional Gaussian specific prior, and stationarity equations

$$\mathbf{K}(\phi - t) = \mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N - \mathbf{P}'(\phi)\Lambda_X, \tag{5.55}$$

$$\left\langle \phi - t \left| \frac{\partial \mathbf{K}(\theta)}{\partial \theta} \right| \phi - t \right\rangle = \mathrm{Tr}\left(\mathbf{K}^{-1}(\theta)\frac{\partial \mathbf{K}(\theta)}{\partial \theta}\right) - 2E'(\theta), \tag{5.56}$$

using Eq. (5.25). In case of a free variation of the matrix elements of the real symmetric, positive (semi–)definite $\mathbf{K}$, the hyperparameters become $\theta_l = \theta(x,y;x',y') = \mathbf{K}(x,y;x',y')$ (which also represent a hyperfield as discussed in detail in Section 5.6). The constraints on $\mathbf{K}$ have to be implemented by the hyperprior $E(\theta)$ on the operator θ if necessary. Then, using

$$\frac{\delta \mathbf{K}(x,y;x',y')}{\delta\theta(x'',y'';x''',y''')} = \delta(x-x'')\delta(y-y'')\delta(x'-x''')\delta(y'-y'''), \tag{5.57}$$

Eq. (5.56) becomes an inhomogeneous equation for the operator θ

$$|\phi - t\rangle\langle\phi - t| = \theta^{-1} - 2E'(\theta). \tag{5.58}$$

To prevent a diverging $\theta = \mathbf{K}$, a nonuniform hyperprior $E(\theta)$ is necessary.

We will in the sequel consider two special cases, one where the determinant of the covariance is θ–independent so that the trace term vanishes (Section 5.4.4), and one where θ is just a multiplicative factor for the specific prior energy (Section 5.4.2).

5.4.2 Automatic relevance detection

A useful application of hyperparameters is the identification of relevant directions within the space of x– and y–variables. Consider the general case of an inverse covariance decomposed into components, $\mathbf{K}_0 = \sum_i \theta_i \mathbf{K}_i$. Treating the

coefficient vector θ (with components θ_i) as hyperparameter with hyperprior $p(\theta)$ results in a prior energy (error) functional

$$\frac{1}{2}\left\langle \phi - t \left| \sum_i \theta_i \mathbf{K}_i \right| \phi - t \right\rangle - \ln p(\theta) + \ln Z_\phi(\theta). \tag{5.59}$$

The θ–dependent term $\ln Z_\phi(\theta)$, resulting from the normalization over ϕ, has to be included to obtain the correct stationarity condition for θ. For a negative Laplacian inverse covariance, for example, the components $\mathbf{K}_i$ are $\mathbf{K}_i = -\partial_{x_i}^2$ or $\mathbf{K}_i = -\partial_{y_i}^2$. In that case adapting the hyperparameters means searching for relevant directions in the space of x– or y–variables. This technique has been called *Automatic Relevance Determination* by MacKay and Neal [319, 320, 363]. The positivity constraint for θ_i can be implemented explicitly, for example, by using $\mathbf{K}_0 = \sum_i \theta_i^2 \mathbf{K}_i$ or $\mathbf{K}_0 = \sum_i \exp(\theta_i)\mathbf{K}_i$.

5.4.3 Local masses and gauge theories

The Bayesian analog of a mass term in field theory is a term proportional to the identity matrix $\mathbf{I}$ in the inverse prior covariance $\mathbf{K}$. Consider,

$$\mathbf{K} = \theta^2 \mathbf{I} - \Delta, \tag{5.60}$$

with positive real θ^2 representing a mass parameter. Such a $\mathbf{K}$ can, for example, be used in Eq. (5.53). For large masses θ^2, the MAP solution ϕ tends, at locations without data points, to be a copy of the prior template t, and longer range effects of training data in Eq. (5.53) resulting from the smoothness induced by Δ become less important. To allow a local adaption of the interplay between 'template copying' and smoothness related reconstruction of training data, the constant mass can be replaced by a function $\theta(z)$, where $z = (x, y)$ for density estimation and $z = x$ for regression. Function hyperparameters like $\theta(z)$ represent additional fields entering the problem, and will be called hyperfields. (Hyperfields will be discussed in more detail in Section 5.6.) Hence going from a fixed hyperparameter θ (mass θ^2) to a function $\theta(z)$ the mass term $\theta^2(\phi(z) - t(z))^2$ takes the form of a ('four point') interaction term $\theta(z)\theta(z)(\phi(z) - t(z))^2$ between the 'relative' field $\phi(z) - t(z)$ and hyperfield $\theta(z)$.

As hyperprior one can use a smoothness constraint on the function $\theta(z)$, analogous to a smoothness prior on $\phi(z)$. For example, choosing a Gaussian smoothness hyperprior with an inverse covariance $\mathbf{K}_\theta$ and mean function $t_\theta(z)$, results in a prior energy

$$\frac{1}{2}\langle \phi - t \,|\mathbf{M}^2|\, \phi - t\rangle - \frac{1}{2}\langle \phi - t \,|\Delta|\, \phi - t\rangle + \langle \theta - t_\theta \,|\mathbf{K}_\theta|\, \theta - t_\theta\rangle + \ln Z_\phi(\theta), \tag{5.61}$$

with $\mathbf{M}$ denoting the diagonal operator with diagonal elements $\theta(x)$. In the language of field theory a hyperprior for $\theta(z)$ corresponds to a Lagrangian density for 'free hyperfields', i.e., for vanishing $\phi(z) - t(z)$.

For example, one can construct hyperfields which show similarities to gauge fields in physics. To obtain a gauge theory–like formalism let us define a diagonal operator $\mathbf{M}^2 = \sum_i \mathbf{M}_i^2$ with diagonal elements $\theta_i^2(z)$, decomposed analogous to $\Delta = \sum_i \Delta_i = \sum_i \partial_i^2$. Then the inverse prior covariance

$$\mathbf{K} = \sum_i \left(\mathbf{M}_i^2 - \partial_i^2\right) = \sum_i \left(\mathbf{M}_i + \partial_i\right)\left(\mathbf{M}_i - \partial_i\right) = \sum_i D_i^\dagger D_i \tag{5.62}$$

can be expressed in terms of a 'covariant derivative' $D_i = \partial_i + \theta_i$, if we choose the 'Lorentz gauge' $\sum_i \partial_i \theta_i = 0$. Furthermore, a 'smoothness' hyperprior energy for $\theta_i(z)$ which resembles the free Lagrangian of Abelian gauge fields can be obtained by choosing

$$\frac{1}{2}\left(\sum_i \langle \theta_i | -\Delta | \theta_i \rangle - \left\langle \sum_i \partial_{x_i}\theta_i \middle| \sum_j \partial_{x_j}\theta_j \right\rangle\right) = \frac{1}{4}\sum_{ij} \langle F_{ij} | F_{ij} \rangle \tag{5.63}$$

which can be expressed in terms of a 'field strength tensor' for Abelian fields,

$$F_{ij} = \partial_i \theta_j - \partial_j \theta_i. \tag{5.64}$$

This is the analogue of the Maxwell tensor in electrodynamics. To relate this, as in electrodynamics, to a local $U(1)$ gauge symmetry $\phi \to e^{i\alpha}\phi$ (for some function $\alpha(x)$ and in combination with $\theta_i \to \theta_i + \partial_i \alpha$), one can consider complex functions ϕ, with the restriction that their phase can not be measured. Notice that, due to the interpretation of the prior as product $p(\phi|\theta)p(\theta)$, an additional θ–dependent normalization term $\ln Z_\phi(\theta)$ enters the energy functional. Such a term is not explicit in field theory (with a Lagrangian density corresponding to the prior energy) if one interprets the prior functional directly as $p(\phi, \theta)$.

5.4.4 Invariant determinants

In this section we discuss parameterizations of the inverse covariance of a Gaussian specific prior which leave the determinant invariant. In this case no θ–dependent normalization factors have to be included, which are usually very difficult to calculate. We have to keep in mind, however, that in general a large freedom for $\mathbf{K}(\theta)$ effectively diminishes the influence of the parameterized prior term.

A determinant is invariant under general similarity transformations, i.e., $\det \tilde{\mathbf{K}} = \det \mathbf{K}$ for $\mathbf{K} \to \tilde{\mathbf{K}} = \mathbf{O}\mathbf{K}\mathbf{O}^{-1}$ where $\mathbf{O}$ could be any element of the general linear group. Similarity transformations do not change the eigenvalues, because $\mathbf{K}\psi = \lambda\psi$ implies $\mathbf{O}\mathbf{K}\mathbf{O}^{-1}\mathbf{O}\psi = \lambda\mathbf{O}\psi$. Thus with $\mathbf{K}$ also $\tilde{\mathbf{K}}$ is

positive definite. The additional constraint that $\tilde{\mathbf{K}}$ has to be real–symmetric,

$$\tilde{\mathbf{K}} = \tilde{\mathbf{K}}^T = \tilde{\mathbf{K}}^\dagger, \tag{5.65}$$

requires $\mathbf{O}$ to be real and orthogonal

$$\mathbf{O}^{-1} = \mathbf{O}^T = \mathbf{O}^\dagger. \tag{5.66}$$

Furthermore, as an overall factor of $\mathbf{O}$ does not change $\tilde{\mathbf{K}}$, one can restrict $\mathbf{O}$ to the special orthogonal group $SO(N)$ with $\det \mathbf{O} = 1$. If $\mathbf{K}$ has degenerate eigenvalues, there exist orthogonal transformations with $\mathbf{K} = \tilde{\mathbf{K}}$.

While in one dimension only the identity remains as transformation, the condition of an invariant determinant becomes less restrictive in higher dimensions. Thus especially for large dimension d of $\mathbf{K}$ (infinite for continuous x) there is a great freedom to adapt inverse covariances without the need to calculate normalization factors.

A positive definite $\mathbf{K}$ can be diagonalized by an orthogonal matrix $\mathbf{O}$ with $\det \mathbf{O} = 1$, i.e., $\mathbf{K} = \mathbf{O}\mathbf{D}\mathbf{O}^T$. Parameterizing $\mathbf{O}(\theta)$ the specific prior term becomes

$$\frac{1}{2} \langle \phi - t \,|\mathbf{K}(\theta)|\, \phi - t \rangle = \frac{1}{2} \left\langle \phi - t \,\middle|\mathbf{O}(\theta)\mathbf{D}\mathbf{O}^T(\theta)\middle|\, \phi - t \right\rangle, \tag{5.67}$$

so the stationarity Eq. (5.56) reads

$$\left\langle \phi - t \left| \frac{\partial \mathbf{O}}{\partial \theta} \mathbf{D}\mathbf{O}^T \right| \phi - t \right\rangle = -E'(\theta). \tag{5.68}$$

Matrices $\mathbf{O}$ from $SO(N)$ include rotations and inversion. For a Gaussian specific prior with non–degenerate eigenvalues, solving Eq. (5.68) allows to choose a small variance in the prior for such dimensions along which the likelihood changes rapidly ('relevant dimensions of the problem') and a large variance for directions along which the likelihood only varies slowly ('less relevant' or 'irrelevant dimensions').

There are also transformations which can change eigenvalues, but leave eigenvectors invariant. As an example, consider a diagonal matrix $\mathbf{D}$ with diagonal elements (and eigenvalues) $\lambda_i \neq 0$, so that $\det \mathbf{D} = \prod_i \lambda_i$. Clearly, any permutation of the eigenvalues λ_i leaves the determinant invariant and leaves a positive definite matrix positive definite. Furthermore, one may introduce continuous parameters $\theta_{ij} > 0$ with $i < j$ and transform $\mathbf{D} \to \tilde{\mathbf{D}}$ according to

$$\lambda_i \to \tilde{\lambda}_i = \lambda_i \theta_{ij}, \quad \lambda_j \to \tilde{\lambda}_j = \frac{\lambda_j}{\theta_{ij}}, \tag{5.69}$$

like, $\tilde{\lambda}_1 = \lambda_1 \theta_{12}$, $\tilde{\lambda}_2 = \lambda_2 / \theta_{12}$ or $\tilde{\lambda}_1 = \lambda_1 \theta_{13}$, $\tilde{\lambda}_2 = \lambda_2$, $\tilde{\lambda}_3 = \lambda_3 / \theta_{13}$. This leaves the product $\lambda_i \lambda_j = \tilde{\lambda}_i \tilde{\lambda}_j$, and therefore also the determinant, invariant and

leaves a positive definite matrix positive definite, because $\theta_{ij} > 0$. This can be done with every pair of eigenvalues defining a set of continuous parameters θ_{ij} with $i < j$ leading to

$$\lambda_i \to \tilde{\lambda}_i = \lambda_i \frac{\prod_{j>i} \theta_{ij}}{\prod_{j<i} \theta_{ji}}, \tag{5.70}$$

like $\tilde{\lambda}_1 = \lambda_1 \theta_{12} \theta_{13}$, $\tilde{\lambda}_2 = \lambda_2 \theta_{23}/\theta_{12}$, $\tilde{\lambda}_3 = \lambda_3/(\theta_{13}\theta_{23})$. This also leaves the determinant invariant

$$\det \tilde{\mathbf{D}} = \prod_i \tilde{\lambda}_i = \prod_i \left(\lambda_i \frac{\prod_{j>i} \theta_{ij}}{\prod_{j<i} \theta_{ji}} \right) = \left(\prod_i \lambda_i \right) \frac{\prod_i \prod_{j>i} \theta_{ij}}{\prod_i \prod_{j<i} \theta_{ji}} = \prod_i \lambda_i = \det \mathbf{D}. \tag{5.71}$$

A transformation with parameterization $\theta_i > 0$, also leaving the eigenvectors unchanged, is

$$\tilde{\lambda}_i = \lambda_i \theta_i, \; i \neq i^*; \quad \tilde{\lambda}_{i^*} = \lambda_{i^*} \prod_{i \neq i^*} \theta_i^{-1}, \tag{5.72}$$

like $\tilde{\lambda}_1 = \lambda_1 \theta_1$, $\tilde{\lambda}_2 = \lambda_2 \theta_2$, $\tilde{\lambda}_3 = \lambda_3/(\theta_1 \theta_2)$, if choosing $i^* = 3$.

These techniques can be applied to a general positive definite $\mathbf{K}$ after diagonalizing

$$\mathbf{K} = \mathbf{O}\mathbf{D}\mathbf{O}^T \to \tilde{\mathbf{K}} = \mathbf{O}\tilde{\mathbf{D}}\mathbf{O}^T \Rightarrow \det \mathbf{K} = \det \tilde{\mathbf{K}}. \tag{5.73}$$

As an example, consider the transformations (5.70, 5.72) for which the specific prior term becomes

$$\frac{1}{2} \langle \phi - t | \mathbf{K}(\theta) | \phi - t \rangle = \frac{1}{2} \langle \phi - t | \mathbf{O}\mathbf{D}(\theta)\mathbf{O}^T | \phi - t \rangle . \tag{5.74}$$

The stationarity Eq. (5.56) then reads

$$\frac{1}{2} \left\langle \phi - t \left| \mathbf{O} \frac{\partial \mathbf{D}}{\partial \theta} \mathbf{O}^T \right| \phi - t \right\rangle = -E'(\theta), \tag{5.75}$$

with, for (5.70) and $k < l$,

$$\frac{\partial \mathbf{D}(i,j)}{\partial \theta_{kl}} = \delta(i-j) \left(\delta(k-i) \, \lambda_k \frac{\prod_{l \neq n > k} \theta_{kn}}{\prod_{n<k} \theta_{nk}} + \delta(l-i) \, \lambda_l \frac{\prod_{n>l} \theta_{ln}}{\prod_{k \neq n < l} \theta_{nl}} \right), \tag{5.76}$$

or, for (5.72) and $k \neq i^*$,

$$\frac{\partial \mathbf{D}(i,j)}{\partial \theta_k} = \delta(i-j) \left(\delta(k-i) \, \lambda_k + \delta(i-i^*) \, \lambda_{i^*} \frac{1}{\theta_k \prod_{l \neq i^*} \theta_l} \right). \tag{5.77}$$

If, for example, $\mathbf{K}$ is a translationally invariant operator, it is diagonalized in a basis of plane waves. Then also $\tilde{\mathbf{K}}$ is translationally invariant, but its sensitivity to certain frequencies has changed. The optimal sensitivity pattern is determined by the given stationarity equations.

5.4.5 Regularization parameters

A hyperparameter which is adapted in nearly all applications of regularization approaches is the factor multiplying the regularization term or, in Bayesian terminology, the inverse temperature of the specific prior. This hyperparameter is often simply called the regularization parameter. The following discussion is easily adapted to parametric approaches with priors on the parameters. For neural networks, for instance, parameterized by its weights $\xi = w_i$, the regularization parameter θ can multiply the sum of squared weights, yielding the prior $\theta \sum_i w_i^2$ [323]. (This prior corresponds to the 'weight decay' algorithm [212].) As the occurring integrals for regularization parameters are non–Gaussian, some care is required when using saddle point approximation techniques for these integrals.

Hence let us write the specific inverse prior covariance as $\mathbf{K}(\lambda) = \lambda \mathbf{K}_0$, where the regularization parameter $\lambda \geq 0$ represents a hyperparameter θ. In regression adapting the hyperparameter λ is equivalent to adapting a global standard deviation or 'noise level' σ in the Gaussian likelihood. The prior normalization factor Z_ϕ then depends on λ because for a Gaussian integral $Z_\phi(\lambda) = (\det \frac{\lambda \mathbf{K}_0}{2\pi})^{-1/2}$. This corresponds to a normalization energy $E_N(\lambda) = -\frac{d}{2} \ln \det \frac{\lambda \mathbf{K}_0}{2\pi} = -\frac{d}{2} \mathrm{Tr} \ln \frac{\lambda \mathbf{K}_0}{2\pi}$. In the case of a discrete d–dimensional ϕ–space, the normalization energy becomes, skipping constants, $E_N(\lambda) = -\frac{d}{2} \ln \lambda$ [cf. (5.7)]. Because of

$$\frac{\partial \mathbf{K}}{\partial \lambda} = \mathbf{K}_0, \tag{5.78}$$

we then find for the stationarity equations, analogous to Eqs. (5.31) and (5.32),

$$\lambda \mathbf{K}_0(\phi - t) = \mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N - \mathbf{P}'(\phi)\Lambda_X, \tag{5.79}$$

$$\frac{1}{2} \langle \phi - t | \mathbf{K}_0 | \phi - t \rangle = \frac{d}{2\lambda} - E'(\lambda), \tag{5.80}$$

when working in a discretized space with dimension d.

For compensating hyperprior the right hand side of Eq. (5.80) vanishes, giving thus no stationary point for λ. Using the condition $\lambda \geq 0$, however, one sees that, for positive definite $\mathbf{K}_0$, Eq. (5.79) is minimized for $\lambda = 0$ corresponding to the 'prior–free' case. For example, in the case of Gaussian regression the solution would be the data template $\phi = t_D$ ('δ–catastrophe', cf. also Sections 3.8, 5.3.3, 5.3.4). To get a nontrivial solution for λ a noncompensating hyperparameter energy $E(\lambda)$ must be used so that $\ln Z_\phi + E_N$ is nonuniform [33,52].

The other limiting case is a vanishing $E'(\lambda)$ for which Eq. (5.80) becomes

$$\lambda = \frac{d}{\langle \phi - t | \mathbf{K}_0 | \phi - t \rangle}. \tag{5.81}$$

For $\phi \to t$ one sees that $\lambda \to \infty$. Moreover, in case $P[t]$ represents a normalized probability (density), $\phi = t$ is also a solution of the first stationarity equation (5.79) in the limit $\lambda \to \infty$. Thus for vanishing $E'(\lambda)$ the 'data–free' solution $\phi = t$ is a selfconsistent solution of the stationarity equations (5.79, 5.80).

Fig. 5.1 shows a posterior surface for uniform and for compensating hyperprior for a one–dimensional regression example. The joint MAP solution corresponds to the highest point of the joint posterior over λ, ϕ in those figures.

Alternatively one can try to integrate the λ–integral analytically and perform a MAP for the posterior with the resulting prior non–Gaussian in ϕ. Using, for example, for the 'scaling' parameter λ, restricted to positive values, the common 'noninformative' (and improper, i.e., nonnormalizable) hyperprior[2] $p(\lambda) = 1/\lambda$ [33, 308, 424, 457], we find for the prior on ϕ integrated over λ (assuming a d–dimensional ϕ–space) a multivariate Student–t distribution [61, 68, 130, 150, 182, 457, 474, 526, 530]

$$\begin{aligned} p(\phi) &= \int_0^\infty d\lambda\, p(\phi|\lambda)p(\lambda) = (2\pi)^{-\frac{d}{2}} (\det \mathbf{K}_0)^{\frac{1}{2}} \int_0^\infty d\lambda\, \lambda^{\frac{d}{2}-1} e^{-\lambda \tilde{E}(\phi)} \quad (5.82) \\ &= (2\pi)^{-\frac{d}{2}} (\det \mathbf{K}_0)^{\frac{1}{2}} \frac{\Gamma(\frac{d}{2})}{\tilde{E}^{\frac{d}{2}}(\phi)} = (2\pi)^{-\frac{d}{2}} (\det \mathbf{K}_0)^{\frac{1}{2}} \Gamma(\tfrac{d}{2})\, e^{-\frac{d}{2} \ln \tilde{E}(\phi)}, \end{aligned}$$

with $\tilde{E}(\phi) = E(\phi)/\lambda$. The λ–independent factor of the prior energy in Eq. (5.82) can, for example, be of the form $\tilde{E}(\phi) = \frac{1}{2} \langle \phi - t\, |\mathbf{K}_0|\, \phi - t \rangle + \tilde{E}_{\min}$, and include a ϕ–independent part $\tilde{E}_{\min}$ analogous to $E_{\min}$ in Eq. (3.195). Such a constant term $\tilde{E}_{\min}$, being the minimal value $\tilde{E}$ can take when varying ϕ, appears in $\tilde{E}$, like in Eq. (3.191), if λ multiplies a sum of 'conflicting' quadratic terms, i.e., terms with different minima t_I. After integrating out the hyperparameter λ, the two MAP stationarity equations (5.79) and (5.80) are reduced to one MAP equation for ϕ with non–Gaussian prior (5.82). For $\tilde{E}_{\min} \neq 0$ the MAP equation reads

$$(2\pi)^{-\frac{d}{2}} \Gamma(\tfrac{d}{2}) (\det \mathbf{K}_0)^{\frac{1}{2}} \frac{d}{2} \frac{\mathbf{K}_0(\phi - t)}{\tilde{E}(\phi)^{\frac{d}{2}+1}} = \mathbf{P}'(\phi) \mathbf{P}^{-1}(\phi) N - \mathbf{P}'(\phi) \Lambda_X. \quad (5.83)$$

In the limit $\tilde{E}_{\min} \to 0$, the solution of Eq. (5.83) approaches the 'data–free' solution $\phi \to t$ if its right hand side is finite at $\phi = t$, i.e., if the template t represents a correctly normalized and non–negative likelihood which is non–zero at the measured data. For example, in Gaussian regression with $\tilde{E}_{\min} \neq 0$ and writing t_0 for the prior template, we have

$$0 = \mathbf{K}_D(\phi - t_D) + (2\pi)^{-\frac{d}{2}} \Gamma(\tfrac{d}{2}) (\det \mathbf{K}_0)^{\frac{1}{2}} \frac{d}{2} \frac{\mathbf{K}_0(\phi - t_0)}{\left(\frac{1}{2} \langle \phi - t_0\, |\mathbf{K}_0|\, \phi - t_0 \rangle + \tilde{E}_{\min} \right)^{\frac{d}{2}+1}}, \quad (5.84)$$

[2] For a multivariate (inverse) covariance, a noninformative prior can be formulated as limit of an (inverse) Wishart distribution [61, 150, 424].

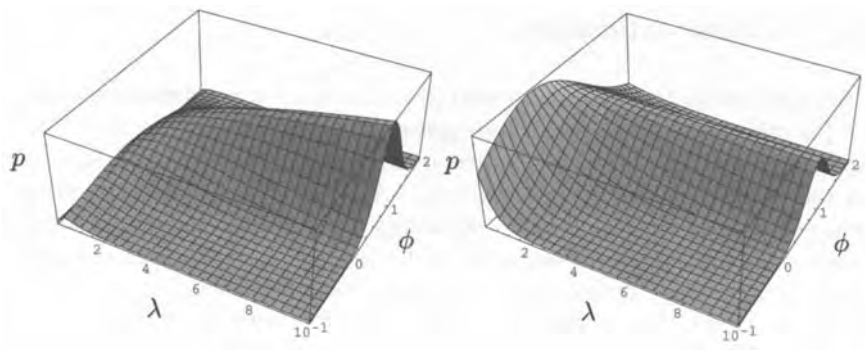


Figure 5.1: Joint posterior density of ϕ and λ, $p(\phi, \lambda | D, D_0) \propto p(y_D|\phi)p(\phi|\lambda, D_0)p(\lambda)$ for an example of Gaussian regression with fixed ('zero–dimensional') x, one–dimensional ϕ–space, and training data $y_D = 0$ and prior data $y_{D_0} = 1$. Left hand side: For uniform prior $p(\lambda) \propto 1$ so that the joint posterior becomes $p \propto e^{-\frac{1}{2}\phi^2 - \frac{\lambda}{2}(\phi-1)^2 + \frac{1}{2}\ln\lambda}$, with maximum at $\lambda = \infty$, $\phi = 1$. Right hand side: For compensating hyperprior $p(\lambda) \propto 1/\sqrt{\lambda}$ so that $p \propto e^{-\frac{1}{2}\phi^2 - \frac{\lambda}{2}(\phi-1)^2}$ with maximum at $\lambda = 0$, $\phi = 0$.

where in the limit $\tilde{E}_{\min} \to 0$ the MAP solution becomes data independent, $\phi \to t_0$. The two possibilities of an evidence approximation and MAP with non–Gaussian prior for ϕ, after integrating out the hyperparameter θ, are compared in [323].

In some cases the θ–integration can also be tackled by Monte Carlo methods [521]. Finally, we remark that specific methods to determine regularization parameters have also been developed in the setting of empirical risk minimization, with a different interpretation of the error functional. Regularization parameters are usually determined by cross–validation [12,89,169,356, 456,471,472,505–507,509], (see also Section 5.4.6), bootstrap [118–120,193], and related techniques, as well as the discrepancy method, self–consistent, or similar techniques [67,89,122,221,222,326,507]. The discrepancy method, which sets (for known standard deviation or 'noise level' σ) $\chi^2 = \sum_{i=1}^n |y_i - \phi(x_i)|^2/\sigma^2 = n$, uses the fact that in regression the regularization parameter plays the same role as the variance in the data term. (For recently developed methods to determine the kernel parameter for support vector machines, which corresponds to a regularization parameter, see [92,440].) Combining a nonparametric Bayesian approach for ϕ with a Frequentist technique like cross–validation to determine the regularization factor λ, seems often to be a good choice in practice.

5.4.6 Cross–validation

A standard method to determine the optimal value of a regularization parameter (or other low–dimensional hyperparameters) is cross–validation. Being a specific plug–in estimate of the prediction error it is a typical Frequentist, and not a genuine Bayesian, technique. As plug–in estimate it requires a sufficient number of data and a sufficiently regular dependency of the prediction error on the hyperparameter under study, but it also has the advantage of being not as model dependent as a typical Bayesian integration over hyperparamaters. We will thus discuss in the following cross–validation for regression [12, 89, 120, 169, 356, 456, 471, 472, 505, 507, 509].

The basic idea of cross–validation is to use only part of the training data to obtain an approximation $\hat{y}(x)$ and use the remaining data as a test set to measure the prediction error. Such plug–in estimates of the prediction error can then be used to select an optimal value for a hyperparameter λ. This is done by calculating from the training data approximations $\hat{y}(\lambda)$ for several values of λ and by using some search algorithm to locate the λ with minimal estimated prediction error.[3]

Splitting the available data into a test and a training set, however, leaves fewer data for training. Using a large test set gives good estimates of the prediction error but not very good training results. A small test set gives good training results but not good estimates of the prediction error. Leave–one–out (or leave–out–one) cross–validation is a clever way to use the available data optimally. To this end all but one data point are used for training, leaving this one point for testing. Then, training is repeated, leaving out another data point which then is used for testing. For n data points this can be repeated n times, yielding n one-point estimates of the prediction error which can then be averaged.[4]

To be specific, let us study a regression problem where we are looking for an approximation $\hat{y}(x)$ for data y at x. We denote by $\hat{y}_i = \hat{y}(x_i)$ the approximation for y_i at x_i resulting from training with the full training data set D, and by $\hat{y}_j^{(-i)} = \hat{y}^{(-i)}(x_j)$ the approximation for y_j at x_j using only the data set D/i which is obtained from D by leaving out the ith data point. In general the approximations $\hat{y}_i(\lambda)$ and $\hat{y}_i^{(-i)}(\lambda)$ depend on λ. The leave–one–

[3] To obtain an estimation of the prediction error for the finally chosen, optimally adapted value of the hyperparameter, a new plug–in estimate of the prediction error has to be calculated on a third independent data set which has not been used for training (training set) or for selecting that optimal value of the hyperparameter (test set).

[4] The price to pay for this optimal use of the available data is the need to do the learning n–times (for every value of λ). Thus instead of leaving one out, in practice one may leave out a certain fraction of the data. Only in the special case of regression, as discussed below in this section, does the linear structure of the problem allow us to avoid learning n times.

out cross–validation error is then defined as

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i^{(-i)}(\lambda) \mid^2 . \tag{5.85}$$

In the nonparametric regression model (3.187) data are approximated by the fields ϕ itself,

$$\hat{y}_i = \phi(x_i). \tag{5.86}$$

In the slightly more general regression model (3.199), corresponding, for example, to classical inverse problems, the fields are indirectly related to observable data. Assuming a linear relationship we have

$$\hat{y}_i = \widetilde{\phi}(x_i) = (\mathbf{F}_D\phi)(x_i) = \int dx\, (\mathbf{F}_D(x_i,x)\phi)(x). \tag{5.87}$$

In the following we will discuss this slightly more general model. The formulas will then also comprise the case of linear regression, interpreting the quantities as indicated in Section 4.4. To calculate the cross–validation error we thus have to obtain

$$y_i - \hat{y}_i^{(-i)} = t_i(x_i) - \widetilde{\phi}^{(-i)}(x_i), \tag{5.88}$$

where $t_i(x_i) = y_i$ and $\widetilde{\phi}^{(-i)}$ is the field obtained from the data leaving out the ith data point. Recall, that for $\widetilde{\phi} = \mathbf{F}_D\,\phi$ with energy

$$\begin{aligned} E(\phi) &= \frac{1}{2}\langle \mathbf{F}_D\phi - t_D \,|\mathbf{K}_D|\, \mathbf{F}_D\phi - t_D\rangle \\ &\quad + \frac{1}{2}\langle \mathbf{F}_0\phi - t_0 \,|\mathbf{K}_0|\, \mathbf{F}_0\phi - t_0\rangle + E^D_{\min} \\ &= \frac{1}{2}\sum_{i=1}^{n}\langle \mathbf{F}_D\phi - t_i \,|\mathbf{K}_i|\, \mathbf{F}_D\phi - t_i\rangle \\ &\quad + \frac{1}{2}\langle \mathbf{F}_0\phi - t_0 \,|\mathbf{K}_0|\, \mathbf{F}_0\phi - t_0\rangle \end{aligned} \tag{5.89}$$

(cf. Eq. (3.198)) we have found in (3.203) that

$$\widetilde{\phi} = \mathbf{F}_D\mathbf{K}^{-1}\left(\mathbf{F}_D^T\mathbf{K}_D t_D + \mathbf{F}_0^T\mathbf{K}_0 t_0\right) = \mathbf{A}\mathbf{K}_D t_D + \mathbf{B}\mathbf{K}_0 t_0, \tag{5.90}$$

with (3.201)

$$\mathbf{K} = \mathbf{F}_D^T\mathbf{K}_D\mathbf{F}_D + \mathbf{F}_0^T\mathbf{K}_0\mathbf{F}_0. \tag{5.91}$$

Analogously, skipping the ith data point results in the energy

$$E^{(-i)}(\phi) = \frac{1}{2}\sum_{\substack{j=1\\ i\neq j}}^{n}\langle \mathbf{F}_D\phi - t_j \,|\mathbf{K}_j|\, \mathbf{F}_D\phi - t_j\rangle + \frac{1}{2}\langle \mathbf{F}_0\phi - t_0 \,|\mathbf{K}_0|\, \mathbf{F}_0\phi - t_0\rangle\,, \tag{5.92}$$

which is minimized by

$$\widetilde{\phi}^{(-i)} = \mathbf{F}_D \left(\mathbf{K}^{(-i)}\right)^{-1} \left(\mathbf{F}_D^T \mathbf{K}_{D/i} t_{D/i} + \mathbf{F}_0^T \mathbf{K}_0 t_0\right), \tag{5.93}$$

where now

$$\mathbf{K}^{(-i)} = \mathbf{F}_D^T \mathbf{K}_{D/i} \mathbf{F}_D + \mathbf{F}_0^T \mathbf{K}_0 \mathbf{F}_0, \tag{5.94}$$

and

$$\mathbf{K}_D = \mathbf{K}_{D/i} + \mathbf{K}_i, \quad \mathbf{K}_D t_D = \mathbf{K}_{D/i} t_{D/i} + \mathbf{K}_i t_i, \quad \mathbf{K} = \mathbf{K}^{(-i)} + \mathbf{K}_i \tag{5.95}$$

because of

$$\mathbf{K}_D = \sum_{i=1}^{n} \mathbf{K}_i, \quad \mathbf{K} = \sum_{i=0}^{n} \mathbf{K}_i, \tag{5.96}$$

and

$$\mathbf{K}_{D/i} = \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbf{K}_j, \quad \mathbf{K}^{(-i)} = \sum_{\substack{j=0 \\ j \neq i}}^{n} \mathbf{K}_j, \quad t_{D/i} = \mathbf{K}_{D/i}^{-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} t_j. \tag{5.97}$$

We want to calculate

$$\begin{aligned} CV &= \frac{1}{n} \sum_{i=1}^{n} \left\langle \mathbf{F}_D \phi^{(-i)} - t_i \middle| \mathbf{K}_i \middle| \mathbf{F}_D \phi^{(-i)} - t_i \right\rangle \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\langle \widetilde{\phi}^{(-i)} - t_i \middle| \mathbf{K}_i \middle| \widetilde{\phi}^{(-i)} - t_i \right\rangle \end{aligned} \tag{5.98}$$

which becomes (5.85) for $\mathbf{K}_i = |x_i\rangle \langle x_i|$, i.e., $\mathbf{K}_i(x, x') = \delta(x - x')\delta(x - x_i)$ and λ–dependent $\widetilde{\phi}^{(-i)}$. To this end we try to express the error terms

$$t_i - \widetilde{\phi}^{(-i)} = t_i - \widetilde{\phi} + \widetilde{\phi} - \widetilde{\phi}^{(-i)} \tag{5.99}$$

in terms of the change of the approximation $\widetilde{\phi}$ if skipping the ith data point. We start by writing

$$\begin{aligned} \hat{y} - \hat{y}^{(-i)} = \widetilde{\phi} - \widetilde{\phi}^{(-i)} &= \mathbf{F}_D \left(\phi - \phi^{(-i)}\right) \\ &= \mathbf{F}_D \Big[\mathbf{K}^{-1} \left(\mathbf{F}_D^T \mathbf{K}_D t_D + \mathbf{F}_0^T \mathbf{K}_0 t_0\right) \\ &\qquad - \left(\mathbf{K}^{(-i)}\right)^{-1} \left(\mathbf{F}_D^T \mathbf{K}_{D/i} t_{D/i} + \mathbf{F}_0^T \mathbf{K}_0 t_0\right) \Big]. \end{aligned} \tag{5.100}$$

Now we split $\left(\mathbf{K}^{(-i)}\right)^{-1}$ into two parts

$$\begin{aligned} \left(\mathbf{K}^{(-i)}\right)^{-1} &= \mathbf{K}^{-1} \mathbf{K} \left(\mathbf{K}^{(-i)}\right)^{-1} \\ &= \mathbf{K}^{-1} \mathbf{F}_D^T \mathbf{K}_D \mathbf{F}_D \left(\mathbf{K}^{(-i)}\right)^{-1} + \mathbf{K}^{-1} \mathbf{F}_0^T \mathbf{K}_0 \mathbf{F}_0 \left(\mathbf{K}^{(-i)}\right)^{-1} \end{aligned} \tag{5.101}$$

using (5.94) and obtain

$$\begin{aligned}\widetilde{\phi} - \widetilde{\phi}^{(-i)} &= \underbrace{\mathbf{F}_D\mathbf{K}^{-1}\left(\mathbf{F}_D^T\mathbf{K}_D t_D + \mathbf{F}_0^T\mathbf{K}_0 t_0\right)}_{\widetilde{\phi}} \\ &\quad - \underbrace{\mathbf{F}_D\mathbf{K}^{-1}\mathbf{F}_D^T}_{\mathbf{A}}\mathbf{K}_D \underbrace{\mathbf{F}_D\left(\mathbf{K}^{(-i)}\right)^{-1}\left(\mathbf{F}_D^T\mathbf{K}_{D/i}t_{D/i} + \mathbf{F}_0^T\mathbf{K}_0 t_0\right)}_{\widetilde{\phi}^{(-i)}} \\ &\quad - \underbrace{\mathbf{F}_D\mathbf{K}^{-1}\mathbf{F}_0^T}_{\mathbf{B}}\mathbf{K}_0\mathbf{F}_0 \underbrace{\left(\mathbf{K}^{(-i)}\right)^{-1}\left(\mathbf{F}_D^T\mathbf{K}_{D/i}t_{D/i} + \mathbf{F}_0^T\mathbf{K}_0 t_0\right)}_{\phi^{(-i)}} \\ &= \widetilde{\phi} - \mathbf{A}\mathbf{K}_D\widetilde{\phi}^{(-i)} - \mathbf{B}\mathbf{K}_0\mathbf{F}_0\phi^{(-i)}. \end{aligned} \tag{5.102}$$

Expressing $\widetilde{\phi}$ according to (5.90) and splitting $\mathbf{K}_D$ as in (5.95) yields

$$\begin{aligned}\widetilde{\phi} - \widetilde{\phi}^{(-i)} &= \mathbf{A}\mathbf{K}_i t_i + \mathbf{A}\mathbf{K}_{D/i}t_{D/i} + \mathbf{F}_D\mathbf{K}^{-1}\mathbf{F}_0^T\mathbf{K}_0 t_0 \\ &\quad -\mathbf{A}\mathbf{K}_i\widetilde{\phi}^{(-i)} - \mathbf{A}\mathbf{K}_{D/i}\widetilde{\phi}^{(-i)} - \mathbf{B}\mathbf{K}_0\mathbf{F}_0\phi^{(-i)} \\ &= \mathbf{A}\mathbf{K}_i\left(t_i - \widetilde{\phi}^{(-i)}\right) + \mathbf{A}\mathbf{K}_{D/i}\left(t_{D/i} - \widetilde{\phi}^{(-i)}\right) \\ &\quad -\mathbf{B}\mathbf{K}_0\left(t_0 - \mathbf{F}_0\phi^{(-i)}\right). \end{aligned} \tag{5.103}$$

Because now according to the MAP–Eq. (3.199) for $\phi^{(-i)}$, multiplied by $\mathbf{F}_D\mathbf{K}^{-1}$,

$$\mathbf{A}\mathbf{K}_{D/i}\left(t_{D/i} - \widetilde{\phi}^{(-i)}\right) = -\mathbf{B}\mathbf{K}_0\left(t_0 - \mathbf{F}_0\phi^{(-i)}\right), \tag{5.104}$$

we finally end up with

$$\widetilde{\phi} - \widetilde{\phi}^{(-i)} = \mathbf{A}\mathbf{K}_i\left(t_i - \widetilde{\phi}^{(-i)}\right). \tag{5.105}$$

This result can be used to eliminate $\widetilde{\phi}^{(-i)}$ from the difference

$$\begin{aligned} t_i - \widetilde{\phi}^{(-i)} &= t_i - \widetilde{\phi} + \widetilde{\phi} - \widetilde{\phi}^{(-i)} \\ &= t_i - \widetilde{\phi} + \mathbf{A}\mathbf{K}_i\left(t_i - \widetilde{\phi}^{(-i)}\right) \\ &= \sum_{k=0}^{\infty}\mathbf{A}\mathbf{K}_i\left(t_i - \widetilde{\phi}\right) \\ &= \left(\mathbf{I} - \mathbf{A}\mathbf{K}_i\right)^{-1}\left(t_i - \widetilde{\phi}\right), \end{aligned} \tag{5.106}$$

so the cross–validation error becomes

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}^{(-i)} - t_i \middle| \mathbf{K}_i \middle| \widetilde{\phi}^{(-i)} - t_i \right\rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}-t_i \left| \left(\mathbf{I}-\mathbf{A}\mathbf{K}_i\right)^{T^{-1}}\mathbf{K}_i\left(\mathbf{I}-\mathbf{A}\mathbf{K}_i\right)^{-1} \right| \widetilde{\phi}-t_i \right\rangle . \quad (5.107)$$

For $\mathbf{K}_i = |x_i\rangle\langle x_i|$, i.e., $\mathbf{K}_i(x,x') = \delta(x-x')\delta(x-x_i)$, this means

$$\widetilde{\phi} - \widetilde{\phi}^{(-i)} = \langle\cdot|\mathbf{A}|x_i\rangle \left\langle x_i \left| y_i - \widetilde{\phi}_i^{(-i)} \right.\right\rangle = \mathbf{A}_{\cdot i}(y_i - \widetilde{\phi}_i^{(-i)}), \quad (5.108)$$

or, evaluated at the data points,

$$\widetilde{\phi}(x_i) - \widetilde{\phi}^{(-i)}(x_i) = \mathbf{A}_{ii}(y_i - \widetilde{\phi}_i^{(-i)}). \quad (5.109)$$

Eq. (5.106) becomes

$$\begin{aligned} y_i - \hat{y}_i^{(-i)} &= y_i - \hat{y}_i + \left(\hat{y} - \hat{y}_i^{(-i)}\right) \\ &= y_i - \hat{y}_i + \mathbf{A}_{ii}\left(y_i - \hat{y}_i^{(-i)}\right) \\ &= (y_i - \hat{y}_i)\sum_{k=0}^{\infty}\mathbf{A}_{ii}^k \\ &= \frac{y_i - \hat{y}_i}{1-\mathbf{A}_{ii}}, \end{aligned} \quad (5.110)$$

and the cross–validation error reads

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i^{(-i)}(\lambda) \mid^2 = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i(\lambda)}{1-\mathbf{A}_{ii}}\right|^2 . \quad (5.111)$$

Generalized cross–validation is obtained by replacing each $\mathbf{A}_{ii}$ in the denominator by its average value $\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_{ii}$ (resulting in certain additional invariance properties [507]),

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{1-\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_{ii}}\right|^2 . \quad (5.112)$$

For $\hat{y} = \widetilde{\phi}$ according to (5.90) this can be written

$$GCV = \frac{1}{n}\frac{\| \left(\mathbf{I}_D - \mathbf{A}\mathbf{K}_D\right) t_D + \mathbf{B}\mathbf{K}_0 t_0 \|^2_{\mathbf{K}_D} + E^D_{\min}}{\left|\frac{1}{\tilde{n}}\mathrm{Tr}\left(\mathbf{I}_D - \mathbf{A}\mathbf{K}_D\right)\right|^2}, \quad (5.113)$$

where $E^D_{\min}$ vanishes and $\tilde{n} = n$ in case all x_i are different and $\mathrm{Tr}\left(\mathbf{A}\mathbf{K}_D\right)$ can be interpreted as effective degrees of freedom for the signal [507]. Much work has been devoted to the numerical evaluation of generalized cross–validation errors (5.112), including large scale problems (see, for example, [70,171]). For general nonlinear problems the cross–validation error (5.85) can not be written in a closed form like (5.111) or (5.112) but must be evaluated by repeating the training process, leaving out each time one (or a certain fraction) of the data points.

5.5 Integer hyperparameters

The hyperparameters θ considered up to now have been real numbers or vectors of real numbers. Such hyperparameters can describe continuous transformations like translation, rotation, or scaling of template functions, and the scaling of inverse covariance operators. For real θ and differentiable posterior, stationarity conditions can be found by differentiating the posterior with respect to θ.

Instead of a class of continuous transformations, a finite number of alternative template functions or inverse covariances may be given. For example, an image to be reconstructed might be expected to show a digit between zero and nine, a letter from some alphabet, or the face of someone who is a member of a known group of people. Similarly, a particular time series may be expected to be either in a high or in a low variance regime. In all these cases, there exist a finite number of classes i which can be represented by specific templates t_i or inverse covariances $\mathbf{K}_i$. Such 'class' variables i are nothing else than hyperparameters θ with integer values. In general, different values of an integer hyperparameter represent different statistical models. For instance, they allow for the comparison of a neural network likelihood model with a decision tree likelihood model. Hence, choosing the optimal value of an integer hyperparameter corresponds to *model selection.*

Binary parameters allow one to select from two reference functions or two inverse covariances that one which fits the data best. Indeed, with

$$t(\theta) = (1-\theta)t_1 + \theta t_2, \tag{5.114}$$

$$\mathbf{K}(\theta) = (1-\theta)\mathbf{K}_1 + \theta\mathbf{K}_2, \tag{5.115}$$

a binary $\theta \in \{0,1\}$ implements hard switching between alternative templates and/or inverse covariances. Since for binary θ, $\theta^2 = \theta$, $(1-\theta)^2 = (1-\theta)$, this corresponds to a conditional prior

$$p(\phi|\theta) \propto e^{-(1-\theta)E_1(\phi) - \theta E_2(\phi)} \tag{5.116}$$

with

$$E_1(\phi) = \frac{1}{2}\left\langle \phi - t_1 \left| \mathbf{K}_1 \right| \phi - t_1 \right\rangle, \tag{5.117}$$

$$E_2(\phi) = \frac{1}{2}\left\langle \phi - t_2 \left| \mathbf{K}_2 \right| \phi - t_2 \right\rangle. \tag{5.118}$$

For integer θ the integral $\int d\theta$ becomes a sum $\sum_\theta$ (we also use the letter i and write $\sum_i$ for integer hyperparameters), so that prior, posterior, and predictive density have the form of a *finite mixture* with components labeled by θ.

Similarly, a real $\theta \in [0,1]$ in (5.114) or (5.115) yields soft mixing. In that case, however, the mixing of templates in (5.114) is not equivalent to a mixing

of prior energies as in (5.116) because for real θ Eqs. (5.114) and (5.115) lead to mixed terms, like $(1-\theta)\theta\,\langle\phi - t_1\,|\mathbf{K}|\,\phi - t_2\rangle\,/2$ for $\mathbf{K}_1 = \mathbf{K}_2$.

For a moderate number of components one may be able to include all of the mixture components in the calculations. If the number of mixture components is too large one must select some of the components, for example, by creating a random sample using Monte Carlo methods or by solving for the θ^* with maximal posterior. In contrast to typical optimization problems for real variables, the corresponding integer optimization problems are usually not very smooth with respect to θ (with smoothness defined in terms of differences instead of derivatives), and are therefore often much harder to solve.

There exists a variety of deterministic and stochastic integer optimization algorithms, which may be combined with ensemble methods like genetic algorithms [99, 167, 218, 274, 339, 346, 444], and with homotopy methods, like simulated annealing [1, 98, 149, 261, 338, 415, 428, 532–534]. Annealing methods are similar to (Markov chain) Monte Carlo methods, which aim at sampling many points from a specific distribution (for example, at fixed temperature). For Monte Carlo methods it is important to have (nearly) independent samples and the correct limiting distribution of the Markov chain. For annealing methods the aim is to find the correct minimum (the ground state having zero temperature) by smoothly changing the temperature from a finite value to zero, but it is less important to model the distribution for nonzero temperatures exactly, although an adequate cooling scheme for lowering the temperature should be used.

Instead of an integer optimization problem one can also try to solve a similar problem for real θ. For example, the binary $\theta \in \{0, 1\}$ in Eqs. (5.114) and (5.115) may be extended to real $\theta \in [0, 1]$. By smoothly increasing an appropriate additional hyperprior $p(\theta)$ one can finally return again to binary hyperparameters $\theta \in \{0, 1\}$.

5.6 Hyperfields

Most hyperparameters θ considered so far have been real or integer *numbers*, or *vectors* with real or integer components θ_i. In this section we will discuss priors parameterized by functions, called *hyperfields*, resulting in a still larger flexibility of the formalism. In a field theoretical language, a hyperfield represents a new Euclidean scalar field which interacts with the field ϕ. In numerical calculations where functions have to be discretized, hyperfields stand for high dimensional hyperparameter vectors.

Using hyperfields one has to keep in mind that a gain in flexibility at the same time tends to lower the influence of the prior. For example, consider as hyperfield a completely adaptive reference function $\theta(x) = t(x)$ within a Gaussian prior (3.156). The template $t(x)$ can stand for a reference for a regression function $\phi(x)$ or, in inverse quantum theory, for a reference $t(x) =$

$v_0(x)$ for a potential $v(x)$. Generalization to a density estimation template $t(x,y)$ for a function $\phi(x,y)$ is straightforward. Then for any $\phi(x)$ the prior energy vanishes for $t(x) = \phi(x)$. In the absence of additional hyperpriors $p(\theta)$ the corresponding MAP solution for the hyperfield $\theta(x) = t(x)$ is thus $\theta^*(x) = \phi(x)$ for which the Gaussian prior (3.156) becomes uniform in ϕ. Hence the price to be paid for the additional flexibility introduced by hyperfields are weaker priors and a large number of additional degrees of freedom. This can considerably complicate calculations and requires sufficiently restrictive hyperpriors for the hyperfields.

Let us define *local hyperfields* $\theta(x)$ to be hyperfields depending on the visible variables x. (In general, hyperfields can be introduced which depend on x and y or on other real variables.) Local hyperfields can be used, for example, to adapt templates or inverse covariances locally. To this end, we express real-symmetric, positive (semi-)definite inverse covariances by square roots or (real) *filter operators* $\mathbf{W}$, so that

$$\mathbf{K} = \mathbf{W}^T\mathbf{W}. \tag{5.119}$$

In components

$$\mathbf{K}(x,x') = \int dx''\, \mathbf{W}^T(x,x'')\mathbf{W}(x'',x'), \tag{5.120}$$

and therefore

$$\begin{aligned} \langle \phi - t\,|\mathbf{K}|\,\phi - t\rangle &= \int dx\,dx'\,dx''\,[\phi(x) - t(x)]\mathbf{W}^T(x,x') \\ &\quad \times\ \mathbf{W}(x',x'')[\phi(x'') - t(x'')] \\ &= \int dx\,|\omega(x)|^2, \end{aligned} \tag{5.121}$$

where we define the *filtered difference*

$$\omega(x) = \int dx'\,\mathbf{W}(x,x')[\phi(x') - t(x')]. \tag{5.122}$$

For instance, a square root (3.136) of the discrete negative Laplacian (3.135) corresponds, for $t \equiv 0$, to a filtered difference $\omega(x) = \phi(x+1) - \phi(x)$.

The exponent of a Gaussian prior for a field ϕ can thus be written as an integral over x,

$$p(\phi) \propto e^{-E(\phi)};\quad E(\phi) = \frac{1}{2}\int dx\,|\omega(x)|^2. \tag{5.123}$$

The representation (5.123) is well suited for introducing local hyperfields. For instance, an adaptive prior

$$p(\phi|\theta) = e^{-E(\phi|\theta)}, \tag{5.124}$$

with a real local hyperfield $\theta(x) \in [0,1]$ can be obtained by mixing locally two alternative filtered differences

$$\omega(x;\theta) = [1-\theta(x)]\,\omega_1(x) + \theta(x)\,\omega_2(x), \tag{5.125}$$

where the two ω_i may differ in their filters and/or reference functions. In that case the hyperfield $\theta(x)$ can locally select the best mixture of the filtered differences ω_i, i.e., that one which yields in (5.124) the largest prior, or smallest prior energy

$$\begin{aligned} E(\phi|\theta) &= \frac{1}{2}\int dx|\omega(x;\theta)|^2 + \ln Z_\phi(\theta) \qquad (5.126)\\ &= \frac{1}{2}\int dx\Big|[1-\theta(x)]\omega_1(x) + \theta(x)\omega_2(x)\Big|^2 + \ln Z_\phi(\theta). \end{aligned}$$

Here the normalization factor

$$Z_\phi(\theta) = \int d\phi\, e^{-\frac{1}{2}\int dx|\omega(x;\theta)|^2} \tag{5.127}$$

depends in general on θ, if the filters of the ω_i differ. Clearly, allowing an unbounded $-\infty \le \theta(x) \le \infty$, any function $\omega(x;\theta)$ can be written in the form of Eq. (5.125), provided $\omega_1(x) \neq \omega_2(x)$ for all x. (We may remark that a real hyperfield $\theta(x)$ is technically similar to the metric introduced in [394, 395] to achieve a reparameterization invariant formulation of the posterior. The metric, however, appears in the likelihood as well and has a different transformation behavior.)

In contrast to soft mixing with real functions $\theta(x)$ a binary local hyperfield $\theta(x) \in \{0,1\}$ implements hard switching between alternative filtered differences. Since in the binary case $\theta^2 = \theta$, $(1-\theta)^2 = (1-\theta)$, and $\theta(1-\theta) = 0$, Eq. (5.126) becomes (cf. Eq. (5.116))

$$E(\phi|\theta) = \frac{1}{2}\int dx\left([1-\theta(x)]|\omega_1(x)|^2 + \theta(x)|\omega_2(x)|^2\right) + \ln Z_\phi(\theta), \tag{5.128}$$

while for real $\theta(x)$ Eq. (5.126) includes a mixed term in $\omega_1\omega_2$. It is sometimes helpful to transform an unrestricted real hyperfield $-\infty \le \widetilde{\theta}(x) \le \infty$ into a bounded real hyperfield $\theta(x) \in [0,1]$ by

$$\theta(x) = \sigma(\widetilde{\theta}(x) - \vartheta), \tag{5.129}$$

with threshold ϑ and sigmoidal transformation

$$\sigma(x) = \frac{1}{1+e^{-2\nu x}} = \frac{1}{2}(\tanh(\nu x) + 1). \tag{5.130}$$

In the limit $\nu \to \infty$ the transformation $\sigma(x)$ of (5.130) approaches the step function $\Theta(x)$ and (5.129) results in a binary $\theta(x) = \Theta(\widetilde{\theta}(x) - \vartheta) \in \{0,1\}$.

Analogous to the global mixing or global switching in Eq. (5.114) and Eq. (5.115), a filtered difference $\omega(x;\theta)$ as in Eq. (5.125) can be constructed by local mixing or switching between template functions $t_1(x')$, $t_2(x')$ or filters $\mathbf{W}_1(x,x')$, $\mathbf{W}_2(x,x')$ using a local hyperfield $\theta(x)$,

$$t_x(x';\theta) = [1-\theta(x)]\,t_1(x') + \theta(x)\,t_2(x'), \tag{5.131}$$

$$\mathbf{W}(x,x';\theta) = [1-\theta(x)]\mathbf{W}_1(x,x') + \theta(x)\mathbf{W}_2(x,x'). \tag{5.132}$$

It is important to note that the local templates or reference functions $t_x(x';\theta)$ are functions of x' and x. Indeed, to obtain a filtered difference $\omega(x;\theta)$ at position x, a reference function t_x is needed for all x' for which the corresponding $\mathbf{W}(x,x')$ is nonzero, since

$$\omega(x;\theta) = \int dx'\,\mathbf{W}(x,x')[\phi(x') - t_x(x';\theta)]. \tag{5.133}$$

In this way the whole template function $t_x(x';\theta)$, rather than individual function values $t(x,\theta)$, is adapted individually for every local filtered difference. In particular, the local reference functions of Eq. (5.131) have to be distinguished from one global, locally adapted reference function

$$t(x';\theta) = [1-\theta(x')]\,t_1(x') + \theta(x')\,t_2(x'), \tag{5.134}$$

which at first glance seems to be the natural generalization of Eq. (5.114) to local hyperfields. Only in Gaussian prior terms with the identity $\mathbf{I}$ as covariance, local template functions $t_x(x',\theta)$ are not required. In that case $t_x(x';\theta)$ is only needed for $x = x'$ and we may directly write $t_x(x';\theta) = \tilde{t}(x';\theta)$, skipping the variable x, and obtain the prior energy

$$\frac{1}{2}\int dx\,|\omega(x;\theta)|^2 = \frac{1}{2}\left\langle \phi - \tilde{t}(\theta)\right|\left.\phi - \tilde{t}(\theta)\right\rangle. \tag{5.135}$$

We remark that one can also generalize Eq. (5.131), which uses the same $t_1(x')$, $t_2(x')$ for all x, by working with reference functions $t_{1,x}(x')$, $t_{2,x}(x')$ which vary with the position x at which the filtered difference $\omega(x)$ is required. This yields

$$t_x(x';\theta) = [1-\theta(x)]\,t_{1,x}(x') + \theta(x)\,t_{2,x}(x'). \tag{5.136}$$

For binary $\theta(x)$ Eq. (5.132) corresponds to an inverse covariance

$$\begin{aligned}\mathbf{K}(\theta) &= \int dx\,\mathbf{K}_x(\theta) = \int dx\,W_x(\theta)W_x^T(\theta)\\ &= \int dx\,\left([1-\theta(x)]W_{1,x}W_{1,x}^T + \theta(x)W_{2,x}W_{2,x}^T\right),\end{aligned} \tag{5.137}$$

with

$$\mathbf{K}_x(\theta) = W_x(\theta)W_x^T(\theta) \tag{5.138}$$

written as dyadic product of the vector $W_x(\theta) = \mathbf{W}(x, \cdot\,; \theta)$ composed out of the analogously defined $W_{i,x} = \mathbf{W}_i(x, \cdot)$. (Note the difference to the notation in Eq. (5.119).) For θ–dependent inverse covariances the normalization factors $Z_\phi(\theta)$ become θ–dependent. They have to be included when integrating over θ or solving for the optimal θ in MAP.

In Eqs. (5.131) and (5.132) it is straightforward to introduce two binary hyperfields θ, θ', one for the reference function t_x and one for the filter $\mathbf{W}$. This results in a conditional prior

$$\begin{aligned} p(\phi|\theta,\theta') &\propto e^{-\frac{1}{2}\int dx \left\langle \phi - t_x(\theta) \middle| \mathbf{K}_x(\theta') \middle| \phi - t_x(\theta) \right\rangle} \\ &= e^{-\frac{1}{2}\int dx\, |\omega(x;\theta,\theta')|^2}. \end{aligned} \tag{5.139}$$

Here we can write

$$\begin{aligned} \int dx\, |\omega(x;\theta,\theta')|^2 &= \langle \phi - t(\theta,\theta') \,|\mathbf{K}(\theta')|\, \phi - t(\theta,\theta') \rangle \\ &\quad + \int dx\, \langle t_x(\theta) \,|\mathbf{K}_x(\theta')|\, t_x(\theta) \rangle \\ &\quad - \langle t(\theta,\theta') \,|\mathbf{K}(\theta')|\, t(\theta,\theta') \rangle, \end{aligned} \tag{5.140}$$

with an effective template $t(\theta,\theta')$ given by

$$t(\theta,\theta') = \mathbf{K}(\theta')^{-1} \int dx\, \mathbf{K}_x(\theta')\, t_x(\theta), \tag{5.141}$$

and effective inverse covariance $\mathbf{K}(\theta') = \int dx\, \mathbf{K}_x(\theta')$ as in Eq. (5.137). Since the last two terms in Eq. (5.140) are ϕ–independent constants (depending only on θ, θ') we see that for fixed hyperfields this prior is minimized by $\phi = t(\theta,\theta')$. For given hyperparameters θ, θ' we can write $p(\phi|\theta,\theta') \propto e^{-E(\phi|\theta,\theta')}$, with a prior energy of the form $E(\phi|\theta,\theta') = \frac{1}{2}\langle \phi - t(\theta,\theta') \,|\mathbf{K}(\theta')|\, \phi - t(\theta,\theta') \rangle$.

As the product of Gaussians is again a Gaussian, several Gaussian prior factors can easily be combined. In this way one can implement a nonlocal property like smoothness and still avoid local template functions $t_x(x',\theta)$ by combining a Gaussian prior with $\mathbf{K} = \mathbf{I}$ as in (5.135) with a Gaussian prior with nondiagonal inverse covariance $\mathbf{K}_2$ and zero (or fixed) template,

$$E(\phi|\theta) = \frac{1}{2}\langle \phi - \tilde{t}(\theta) | \, \phi - \tilde{t}(\theta) \rangle + \frac{1}{2}\langle \phi \,|\mathbf{K}_2|\, \phi \rangle. \tag{5.142}$$

Combining both terms yields

$$E(\phi|\theta) = \frac{1}{2}\Big(\langle \phi - t(\theta) \,|\mathbf{K}|\, \phi - t(\theta) \rangle + \langle \tilde{t}(\theta) \,|\mathbf{I} - \mathbf{K}^{-1}| \, \tilde{t}(\theta) \rangle \Big), \tag{5.143}$$

with the second term being independent of ϕ and with effective template and effective inverse covariance

$$t(\theta) = \mathbf{K}^{-1}\tilde{t}(\theta), \quad \mathbf{K} = \mathbf{I} + \mathbf{K}_2. \tag{5.144}$$

For differential operators $\mathbf{K}$ the effective $t(\theta)$ is thus a smoothed version of $\tilde{t}(\theta)$.

The extreme case would be to treat t and $\mathbf{W}$ themselves as unrestricted hyperfields, which just eliminates the corresponding prior term. To restrict the flexibility, a smoothness hyperprior may be imposed to prevent highly oscillating functions $\theta(x)$. For real $\theta(x)$, for example, a smoothness prior like a Laplacian prior $\langle\theta\,|{-}\Delta|\,\theta\rangle\,/2$ can be used in regions where it is defined. (The space of functions for which a smoothness prior with discontinuous templates is defined depends on the locations of the discontinuities.) An example of a non–Gaussian hyperprior is

$$p(\theta) \propto e^{-\frac{\tau}{2}\int dx\, C_\theta(x)}, \tag{5.145}$$

where τ is a constant and

$$C_\theta(x) = \sigma\left(\left(\frac{\partial\theta}{\partial x}\right)^2 - \vartheta_\theta\right), \tag{5.146}$$

with sigmoid $\sigma(x)$ as in (5.130). For $\nu \to \infty$ the sigmoid approaches a step function and $C_\theta(x)$ becomes 0 at locations where the square of the first derivative is smaller than a certain threshold $0 \leq \vartheta_\theta < \infty$, and 1 otherwise. For discrete x one can analogously count the number of jumps larger than a given threshold. One can then penalize the number $N_d(\theta)$ of discontinuities, where $(\partial\theta/\partial x)^2 = \infty$, using

$$p(\theta) \propto e^{-\frac{\tau}{2}N_d(\theta)}. \tag{5.147}$$

In the case of a binary field this corresponds to counting the number of times the field changes its value. The expression C_θ of Eq. (5.146) can be generalized to

$$C_\theta(x) = \sigma\left(|\omega_\theta(x)|^2 - \vartheta_\theta\right), \tag{5.148}$$

where, analogous to Eq. (5.122),

$$\omega_\theta(x) = \int dx'\, \mathbf{W}_\theta(x,x')[\theta(x') - t_\theta(x')], \tag{5.149}$$

with template $t_\theta(x')$ representing the expected form for the hyperfield, and a filter operator $\mathbf{W}_\theta$ defining a distance measure for hyperfields. Parameters of the hyperprior like τ in Eq. (5.145) or Eq. (5.147) can be treated as higher level hyperparameters.

5.7 Auxiliary fields

As an alternative to introducing hyperfields $\theta(x)$ one can work with priors which are explicitly non–Gaussian with respect to ϕ. This can be done by

introducing auxiliary fields $B(x;\phi)$, or more general $B(x,y;\phi)$, whose function values are not considered as independent variables but are directly defined as functionals of ϕ. (For the sake of simplicity we will for $B(x;\phi)$ also write $B(x)$ or $B(\phi)$, depending on the context.) Like hyperfields, auxiliary fields can select locally the best adapted filtered difference from a set of alternative ω_i, each of the form of Eq. (5.122).

For instance, consider the auxiliary field [cf. Eqs. (5.129) or (5.148)]

$$B(x) = \sigma\left(u(x) - \vartheta\right), \tag{5.150}$$

where

$$u(x) = |\omega_1(x)|^2 - |\omega_2(x)|^2, \tag{5.151}$$

ϑ represents a threshold, $\sigma(x)$ a sigmoidal function as in (5.130), and the ω_i are filtered differences defined in terms of ϕ according to Eq. (5.122). Again a binary field $B(x)$ is obtained by letting the sigmoid approach the step function. Because the ω_i depend on ϕ, it is clear from the definition (5.150) that the auxiliary field $B(x)$ is no independent hyperfield but has values being functionals of ϕ. Notice that $B(x)$ is nonlocal with respect to $\phi(x)$ if $\omega_i(x)$ is nonlocal; a value $B(x)$ then depends on more than one $\phi(x)$–value. For a negative Laplacian prior in one–dimension, Eq. (5.150) reads

$$B(x) = \sigma\left(\left|\frac{\partial(\phi - t_1)}{\partial x}\right|^2 - \left|\frac{\partial(\phi - t_2)}{\partial x}\right|^2 - \vartheta\right). \tag{5.152}$$

While auxiliary fields $B(x)$ are directly determined by ϕ, hyperfields are indirectly coupled to ϕ through the MAP stationarity equations. Conversely, an auxiliary field $B(x)$ can be treated formally as an independent hyperfield if a Lagrange multiplier term $\lambda\left[B(x) - \sigma\left(u(x) - \vartheta\right)\right]$ is added to the prior energy in the limit $\lambda \to \infty$.

Like hyperfields $\theta(x)$, auxiliary fields $B(x)$ can be used to adapt reference functions t or filters $\mathbf{W}$. However, a prior as in Eq. (3.156) is non–Gaussian with respect to ϕ if $t(B)$ and $\mathbf{K}(B)$ depend on B and thus also on ϕ. Furthermore, analogous to hyperpriors $p(\theta)$, additional prior terms $p(B(\phi)) \propto \exp(-E_B(\phi))$ for ϕ can be included, formulated in terms of an auxiliary field $B(x)$. As in Eq. (5.128) a binary $B(x)$ can switch between two filtered differences

$$|\omega(x;B)|^2 = [1 - B(x)]|\omega_1(x)|^2 + B(x)|\omega_2(x)|^2, \tag{5.153}$$

within a (non–Gaussian) prior for ϕ

$$p(\phi) \propto e^{-E(\phi) - E_B(\phi)}, \tag{5.154}$$

where the normalization factor $Z = \int d\phi\, e^{-E(\phi) - E_B(\phi)}$ of (5.154) is by definition independent of ϕ. Hence it can be skipped for MAP calculations even

for non–Gaussian $p(\phi)$. In Eq. (5.154)

$$E(\phi) = \frac{1}{2}\int dx\ \left([1 - B(x)]|\omega_1(x)|^2 + B(x)|\omega_2(x)|^2\right), \tag{5.155}$$

depends on ϕ through $B(\phi)$ and $\omega_i(\phi)$ according to Eq. (5.153), while $E_B(\phi)$ depends on ϕ only through $B(\phi)$. For example, the number of switchings can be restricted by taking

$$E_B(\phi) = \frac{\tau}{2}N_d(B), \tag{5.156}$$

where $N_d(B)$ counts the number of discontinuities of $B(x)$. Other choices, for real $B(x)$, are quadratic energies

$$E_B(\phi) = \frac{\tau}{2}\int dx|\omega_B(x)|^2 \tag{5.157}$$

or non–quadratic energies of the form

$$E_B(\phi) = \frac{\tau}{2}\int dx\, C_B(x), \tag{5.158}$$

where, similar to (5.148),

$$C_B(x) = \sigma\left(|\omega_B(x)|^2 - \vartheta_B\right) \tag{5.159}$$

and

$$\omega_B(x) = \int dx'\, \mathbf{W}_B(x,x')[B(x') - t_B(x')] \tag{5.160}$$

is a filtered difference of B with filter operator $\mathbf{W}_B$ and template t_B.

Let us compare a non–Gaussian prior built of prior energies (5.155) and (5.156) for a binary auxiliary field (5.150)

$$p(\phi) \propto e^{-\frac{1}{2}\int dx\left([1-B(x)]|\omega_1(x)|^2+B(x)|\omega_2(x)|^2\right)-\frac{\tau}{2}N_d(B)}, \tag{5.161}$$

with the similar–looking combination of Gaussian prior (5.128) with hyperprior (5.147) for a binary hyperfield,

$$p(\phi,\theta) = p(\phi|\theta)p(\theta) \propto \tag{5.162}$$

$$e^{-\frac{1}{2}\int dx\left[(1-\theta(x))|\omega_1(x)|^2+\theta(x)|\omega_2(x)|^2\right]-\frac{\tau}{2}N_d(\theta)-\ln Z_\phi(\theta)}.$$

Eq. (5.162) works with conditional priors $p(\phi|\theta)$, hence the corresponding normalization factors are in general θ–dependent and have to be included for MAP calculations. Typically, MAP solutions for B, $N_d(B)$, and C_B, being directly defined in terms of the corresponding MAP solution for ϕ, are different from the MAP solutions for θ, $N_d(\theta)$, and C_θ, respectively. However, if the filtered differences ω_i in Eq. (5.162) differ only in their templates, the

normalization term can be skipped. Then, assuming $\vartheta = 0$, $p(\theta) \propto 1$, $p(B) \propto 1$, the two equations are equivalent for $\theta(x) = \Theta\left(|\omega_1(x)|^2 - |\omega_2(x)|^2\right)$. In the absence of hyperpriors, it is indeed easily seen that this is a selfconsistent solution for θ for every given ϕ. In general, however, when hyperpriors are included, another solution for θ may have a larger posterior.

Hyperpriors $p(\theta)$ or additional auxiliary prior terms $p(B)$ can be useful to enforce specific global constraints for $\theta(x)$ or $B(x)$. In natural images, for example, discontinuities are expected to form closed curves. Priors or hyperpriors, organizing discontinuities along lines or closed curves, are thus important for image segmentation or image restoration [39, 146, 152, 332, 401, 539]. A similar problem appears in the determination of piecewise smooth relaxation time spectra from rheological data [431].

5.8 Non–quadratic potentials

Solving learning problems numerically by discretizing the x– and y–variables allows one, in principle, to deal with arbitrary non–Gaussian priors. Compared to Gaussian priors, however, the resulting stationarity equations are intrinsically nonlinear already for Gaussian regression problems.

As a typical example let us formulate a prior in terms of nonlinear, non–quadratic 'potential' functions $\psi(\omega)$ acting on 'filtered differences' $\omega = \mathbf{W}(\phi - t)$, defined with respect to some positive (semi–)definite inverse covariance $\mathbf{K} = \mathbf{W}^T\mathbf{W}$. In particular, consider a prior factor of the following form

$$p(\phi) = e^{-\int dx\, \psi(\omega(x)) - \ln Z_\phi} = \frac{e^{-E(\phi)}}{Z_\phi}, \tag{5.163}$$

where $E(\phi) = \int dx\, \psi(\omega(x))$. (For general density estimation problems x has to be replaced by the pair (x, y).) Non–Gaussian priors are, for example, used to deal with discontinuities in images [57, 151, 152, 359, 537, 538] or in the identification of 'outliers' like the separation of background and signal in experimental spectra [112, 130, 499].

For differentiable ψ function, the functional derivative with respect to $\phi(x)$ becomes

$$\delta_{\phi(x)} p(\phi) = -e^{-\int dx'\, \psi(\omega(x')) - \ln Z_\phi} \int dx''\, \psi'(\omega(x''))\mathbf{W}(x'', x), \tag{5.164}$$

with derivative $\psi'(\omega) = d\psi(\omega)/d\omega$ (and coordinates x', x''), or $\delta_\phi p(\phi) = -p(\phi)\, \mathbf{W}^T\psi'$ from which follows

$$\delta_\phi E(\phi) = -\delta_\phi \ln p(\phi) = \mathbf{W}^T\psi'. \tag{5.165}$$

For nonlinear filters acting on $\phi - t$, $\mathbf{W}$ in Eq. (5.164) must be replaced by $\omega'(x) = \delta_{\phi(x)}\omega(x)$. Instead of a single $\mathbf{W}$ a 'filter bank' with several $\mathbf{W}_\alpha$ and

corresponding $\mathbf{K}_\alpha$, ω_α, ψ_α may be used, so that

$$p(\phi) = e^{-\sum_\alpha \int dx\, \psi_\alpha(\omega_\alpha(x)) - \ln Z_\phi}, \tag{5.166}$$

and

$$\delta_\phi E(\phi) = \sum_\alpha \mathbf{W}_\alpha^T \psi_\alpha'. \tag{5.167}$$

The potential functions ψ may be fixed in advance for a given problem. Typical choices to allow discontinuities are symmetric 'cup' functions with minimum at zero and flat tails for which one large step is cheaper than many small ones [524]. Examples are shown in Fig. 5.2 (a,b). The cusp in (b), where the derivative does not exist, requires special treatment [537]. Such functions can also be interpreted in the sense of robust statistics, since flat tails reduce the sensitivity with respect to outliers [55, 147, 177, 196, 213, 224–226, 229, 432, 442, 469, 494, 497]. Flat (or long) tails can, for example, be used to estimate the background in experimental spectra treating the signal as outlier [112, 130, 499].

Inverted 'cup' functions, like those shown in Fig. 5.2 (c), have been obtained by optimizing a set of ψ_α with respect to a sample of natural images [537]. (For statistics of natural images, their relation to wavelet–like filters, and sparse coding, see also [372, 373].)

While for differential operators $\mathbf{W}$, cup functions induce smoothness, inverse cup functions can be used to implement structure. For such $\mathbf{W}$ the gradient algorithm for minimizing $E(\phi)$,

$$\phi_{\text{new}} = \phi_{\text{old}} - \eta \delta_\phi E(\phi_{\text{old}}), \tag{5.168}$$

becomes in the continuum limit a nonlinear parabolic partial differential equation

$$\phi_\tau = -\sum_\alpha \mathbf{W}_\alpha^T \psi_\alpha'(\mathbf{W}_\alpha(\phi - t)). \tag{5.169}$$

Here a formal time variable τ has been introduced so that $(\phi_{\text{new}} - \phi_{\text{old}})/\eta \to \phi_\tau = d\phi/d\tau$. For cup functions this equation is of diffusion type [368, 396], if also inverted cup functions are included the equation is of reaction–diffusion type [537]. Such equations are known to generate a great variety of patterns.

Alternatively to fixing ψ in advance or, which is sometimes possible for low–dimensional discrete function spaces like images, to approximate ψ by sampling from the prior distribution, one may also introduce hyperparameters and adapt potentials $\psi(\theta)$ to the data.

For example, attempting to adapt an unrestricted function $\psi(\omega)$ with hyperprior $p(\psi)$ by MAP, one has to solve the stationarity condition

$$0 = \delta_{\psi(\omega')} \ln p(\phi, \psi) = \delta_{\psi(\omega')} \ln p(\phi|\psi) + \delta_{\psi(\omega')} \ln p(\psi), \tag{5.170}$$

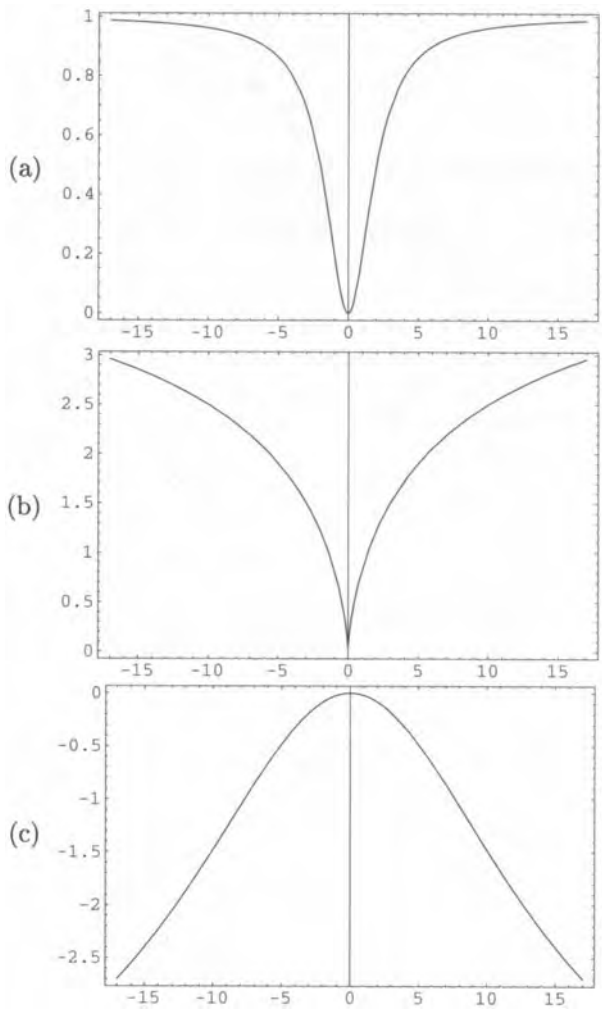


Figure 5.2: Non–quadratic potentials of the form $\psi(x) = a(1.0 - 1/(1 + (|x - x_0|/b)^\gamma))$, [537]. 'Diffusion terms': (a) Winkler's cup function [524] ($a = 5$, $b = 10$, $\gamma = 0.7$, $x_0 = 0$), (b) with cusp ($a = 1$, $b = 3$, $\gamma = 2$, $x_0 = 0$), (c) 'Reaction term' ($a = -4.8$, $b = 15$, $\gamma = 2.0$, $x_0 = 0$).

for $p(\phi|\psi) = e^{-E(\phi|\psi)}/Z_\phi(\psi)$ of the form of (5.163). From

$$\begin{aligned}\delta_{\psi(\omega')}p(\phi|\psi) &= -p(\phi|\psi)\int dx\,\delta\left(\omega' - \omega(x)\right) - \frac{e^{-E(\phi|\psi)}}{Z_\phi^2}\delta_{\psi(\omega')}Z_\phi \\ &= -p(\phi|\psi)\left(n(\omega',\phi) + \int d\phi' p(\phi'|\psi)\, n(\omega',\phi')\right), \quad (5.171)\end{aligned}$$

with integer

$$n(\omega',\phi) = \int dx\,\delta\left(\omega' - \omega(x,\phi)\right), \quad (5.172)$$

being the histogram of the filtered differences, it follows for the first term on the right hand side of Eq. (5.170)

$$-\delta_{\psi(\omega')}\ln p(\phi|\psi) = \delta_{\psi(\omega')}\left[E(\phi|\psi) + \ln Z_\phi(\psi)\right] = n(\omega') - <n(\omega')>, \quad (5.173)$$

with average histogram

$$<n(\omega',\phi)> = \int d\phi\, p(\phi|\psi)\, n(\omega',\phi). \quad (5.174)$$

The right hand side of Eq. (5.173) is zero at ϕ^* if, e.g., $p(\phi|\psi) = \delta(\phi - \phi^*)$, which is the case for $\psi(\omega(x;\phi)) = \beta\left(\omega(x;\phi) - \omega(x;\phi^*)\right)^2$ in the $\beta \to \infty$ limit.

The basic prior energy variants discussed in this Section are applied in Section 7.1.5 to problems of inverse quantum statistics and summarized in Table 7.2.

This page intentionally left blank

Chapter 6

Mixtures of Gaussian prior factors

6.1 Multimodal energy surfaces

Non–Gaussian prior factors which correspond to multimodal energy surfaces can be constructed or approximated by using mixtures of simpler prior components. In particular, it is convenient to use Gaussian densities as components or 'building blocks', since then many useful results obtained for Gaussian processes survive the generalization to mixture models [291–295]. We will therefore in the following discuss applications of mixtures of Gaussian priors. Other implementations of non–Gaussian priors were already discussed in Section 5.8.

In Section 5.1 we have seen that hyperparameters label components of mixture densities. Thus if j labels the components of a mixture model, then j can be viewed as hyperparameter. In Section 5 we have treated the corresponding hyperparameter integration completely in saddle point approximation. In this section we will assume the hyperparameters j to be discrete and try to perform the corresponding summation exactly.

Hence consider a discrete hyperparameter j, possibly in addition to continuous hyperparameters θ. In contrast to the θ–integral we now aim at treating the analogous sum over j exactly, i.e., we want to study *mixture models*

$$p(\phi,\theta|\tilde{D}_0) = \sum_j^m p(\phi,\theta,j|\tilde{D}_0) = \sum_j^m p(\phi|\tilde{D}_0,\theta,j)p(\theta,j). \qquad (6.1)$$

In the following we concentrate on *mixtures of Gaussian specific priors*. Notice that Gaussian mixture models for the prior do *not* correspond to Gaussian mixture models for the likelihood ϕ, the latter being of common use in density

estimation. Indeed, under a mixture of Gaussian priors it is only the prior or posterior density which is modeled by a mixture, the form of ϕ is not restricted to a Gaussian mixture but can be completely unrestricted.

We also remark that an asymptotical interpretation of a saddle point approximation in the mixture case is possible in terms of a parameter $\tilde{\beta}$ when writing $p(\phi, \theta|\tilde{D}_0) \propto e^{\tilde{\beta} \ln \sum_j p(\phi,\theta,j|\tilde{D}_0)}$

We already discussed briefly in Section 5.3 that, in contrast to a product of probabilities (densities) or a sum of error terms implementing a probabilistic AND of approximation conditions, the sum over j in Eq. (6.1) implements a probabilistic OR. Alternative approximation conditions can thus be implemented by giving a corresponding set of alternative templates t_j and inverse covariances $\mathbf{K}_j$. A prior (or posterior) density in form of a probabilistic OR means that the optimal solution does not have to approximate all but only one of those t_j, in a metric defined by $\mathbf{K}_j$. Technically speaking, mixtures of Gaussian prior factors allow to construct nonconvex prior energy surfaces. Different local minima representing alternative *a priori* constraints can thus be adapted to the specific problem under consideration. For example, we may expect in an image reconstruction task blue or brown eyes whereas a mixture between blue and brown might not be as likely. Prior mixture models are potentially useful for [291–295]:

1. Ambiguous (prior) data: Alternative templates can, for example, represent different expected trends for a time series.

2. Model selection: Here templates represent alternative reference models (like different neural network architectures or decision trees); determining the optimal θ corresponds to training such models.

3. Expert knowledge: *A priori* knowledge can be formulated (for example, verbally) in terms of conjunctions and disjunctions of simple components or building blocks. Assume an image of a face is expected to contain certain constituents (AND with respect to eyes, mouth, nose) appearing in various possible variants (OR). Representing the simple components/building blocks by Gaussian priors centered around typical prototypes (e.g., typical eyes) results in Gaussian mixture models. This constitutes a possible link between symbolic information processing and statistical methods [11, 50, 127, 178, 205, 230, 433]. Such an application of prior mixture models has some similarities with the quantification of 'linguistic variables' by fuzzy methods [252, 265, 266].

Finally we mention that besides prior mixture models there are also approaches which resemble the form of Landau–Ginzburg models in statistical mechanics that can be used to implement an OR–like combination of prior factors [291]. Prior mixture models, however, have a clearer probabilistic foundation and are technically more convenient.

6.2 Prior mixtures for density estimation

First we discuss the construction of multimodal energy/error functionals for density estimation using the mixture approach (6.1). For Gaussian components Eq. (6.1) results in an energy functional

$$\begin{aligned} E_{\theta,\phi} &= -\langle \ln P(\phi)|\, N\rangle + \langle P(\phi)|\, \Lambda_X\rangle \\ &\quad - \ln \sum_j e^{-\left(\frac{1}{2}\left\langle \phi - t_j(\theta)\middle|\mathbf{K}_j(\theta)\middle|\phi - t_j(\theta)\right\rangle + \ln Z_\phi(\theta,j) + E_{\theta,j}\right)}, \qquad (6.2) \\ &= -\ln \sum_j e^{-E_{\phi,j} - E_{\theta,j} + c_j}, \qquad (6.3) \end{aligned}$$

where

$$E_{\phi,j} = -\langle \ln P(\phi)|\, N\rangle + \langle P(\phi)|\, \Lambda_X\rangle + \frac{1}{2}\langle \phi - t_j(\theta)\,|\mathbf{K}_j(\theta)|\,\phi - t_j(\theta)\rangle\,, \qquad (6.4)$$

writing the arguments of energies in this section as indices, and

$$c_j = -\ln Z_\phi(\theta, j). \qquad (6.5)$$

The stationarity equations for ϕ and θ

$$0 = \sum_j^m \frac{\delta E_{\phi,j}}{\delta \phi}\, e^{-E_{\phi,j} - E_{\theta,j} + c_j}, \qquad (6.6)$$

$$0 = \sum_j^m \left(\frac{\partial E_{\phi,j}}{\partial \theta} + \frac{\partial E_{\theta,j}}{\partial \theta} + \mathbf{Z}'_j Z_\phi^{-1}(\theta, j)\right) e^{-E_{\phi,j} - E_{\theta,j} + c_j}, \qquad (6.7)$$

with $\mathbf{Z}'_j = \frac{\partial Z_\phi(\theta,j)}{\partial \theta}$, for multidimensional θ generalized as in Eq. (5.14), can also be written

$$0 = \sum_j^m \frac{\delta E_{\phi,j}}{\delta \phi}\, p(\phi, \theta, j|\tilde{D}_0), \qquad (6.8)$$

$$0 = \sum_j^m \left(\frac{\partial E_{\phi,j}}{\partial \theta} + \frac{\partial E_{\theta,j}}{\partial \theta} + \mathbf{Z}'_j Z_\phi^{-1}(\theta, j)\right) p(\phi, \theta, j|\tilde{D}_0). \qquad (6.9)$$

Analogous equations are obtained for parameterized $\phi(\xi)$.

6.3 Numerical case study: Prior mixtures for density estimation

Having seen Bayesian field theoretical models working in Section 3.4 for Gaussian prior factors we will study in this section the slightly more complex prior

mixture models. As discussed in Section 6.1, prior mixture models are useful for implementing complex and unsharp prior knowledge. In particular, they provide a tool to translate verbal statements of experts into quantitative prior densities [291–295], similar to the quantification of 'linguistic variables' by fuzzy methods [265, 266].

We will now study a prior mixture with Gaussian prior components in the function $\phi(x,y) = L(x,y)$ representing the log–likelihood $\ln p(y|x,\phi) = \ln p(y|x,L)$. Consider the following energy functional with mixture prior

$$E(L) = -\ln \sum_j p_j e^{-E_j} = -\langle L|\, N\rangle + \langle e^L|\, \Lambda_X\rangle - \ln \sum_j p_j e^{-\lambda E_{0,j}} \tag{6.10}$$

with given mixture probabilities p_j and with mixture components

$$E_j = -\langle L|\, N\rangle + \lambda E_{0,j} + \langle e^L|\, \Lambda_X\rangle\,. \tag{6.11}$$

We choose Gaussian component prior factors with equal covariances but differing means

$$E_{0,j} = \frac{1}{2}\langle L - t_j\,|\mathbf{K}|\, L - t_j\rangle\,. \tag{6.12}$$

Hence the stationarity equation for functional (6.10) becomes

$$0 = N - \lambda \mathbf{K}\left(L - \sum_j a_j t_j\right) - \mathbf{e}^{\mathbf{L}}\Lambda_X, \tag{6.13}$$

with Lagrange multiplier function

$$\Lambda_X = N_X - \lambda \mathbf{I}_X \mathbf{K}\left(L - \sum_j a_j t_j\right) \tag{6.14}$$

and mixture coefficients

$$a_j = \frac{p_j e^{-\lambda E_{0,j}}}{\sum_k p_k e^{-\lambda E_{0,k}}}. \tag{6.15}$$

The parameter λ represents an inverse temperature for the Gaussian mixture prior. (A different inverse temperature β, multiplying also the data terms, will be chosen for regression in Section 6.4.) Thus, in the 'low temperature' limit $\lambda \to \infty$, the prior component with minimal component energy $E_{0,j}$ dominates the functional (6.10). The solutions $L = t_j$ then represent local minima of the energy (6.10), provided the prior templates t_j in Eq. (6.12) are pairwise different and fulfill the positivity and normalization conditions. Similarly for $\lambda \to 0$ all prior components become equal and the prior factor is effectively switched off.

Typical numerical results for a prior mixture model with two mixture components are presented in Figs. 6.2 – 6.5. Like for Fig. 3.5, the true likelihood

used for these calculations is given by Eq. (3.93) and shown in Fig. 6.1. The corresponding true regression function is thus that of Fig. 3.8. Also, the same training data have been used as for the model of Fig. 3.5 (Fig. 3.7). The two templates t_1 and t_2 which have been selected for the two prior mixture components are (Fig. 6.1)

$$t_1(x,y) = \frac{1}{2\sqrt{2\pi}\sigma_t}\left(e^{-\frac{(y-\mu_a)^2}{2\sigma_t^2}} + e^{-\frac{(y-\mu_b)^2}{2\sigma_t^2}}\right), \tag{6.16}$$

$$t_2(x,y) = \frac{1}{\sqrt{2\pi}\sigma_t}e^{-\frac{(y-\mu_2)^2}{2\sigma_t^2}}, \tag{6.17}$$

with $\sigma_t = 2$, $\mu_a = \mu_2 + 25/9 = 10.27$, $\mu_b = \mu_2 - 25/9 = 4.72$, and $\mu_2 = 15/2$. Both templates capture a bit of the structure of the true likelihood, but not too much, hence learning remains interesting. The average test error of t_1 is equal to 2.56 and is thus lower than that of t_2 being equal to 2.90. The minimal possible average test error 2.23 is given by that of the true solution P_{true}. A uniform likelihood $P = \exp L$, being the effective template in the zero mean case of Fig. 3.5, has an average test error 2.68 between the two templates t_1 and t_2.

Fig. 6.2 proves that convergence is fast for massive prior relaxation when starting from t_1 as initial guess $L^{(0)}$. Compared to Fig. 3.5 the solution is a bit smoother, and, as template t_1 is a better reference than the uniform likelihood, the final test error is slightly lower than for the zero–mean Gaussian prior on L. Starting from $L^{(0)} = t_2$ or a uniform initial guess, a similar, but not exactly the same solution, is obtained. The effect of changing the λ parameter of the prior mixture can be seen in Fig. 6.3 and Fig. 6.4. Larger λ yields a smoother solution and faster convergence when starting from a smooth likelihood template (Fig. 6.3). Smaller λ results in a more rugged solution combined with a slower convergence. The test error in Fig. 6.4 already indicates overfitting.

Prior mixture models tend to produce metastable and approximately stable solutions. Fig. 6.5 presents an example where for an initial guess $L^{(0)} = t_2$ the learning algorithm seems to have produced a stable solution after a few iterations. However, iterating long enough, it decays into a solution with smaller distance to t_1 and with lower test error. Notice that this effect can be prevented by starting with another initialization, as with $L^{(0)} = t_1$ or a similar initial guess.

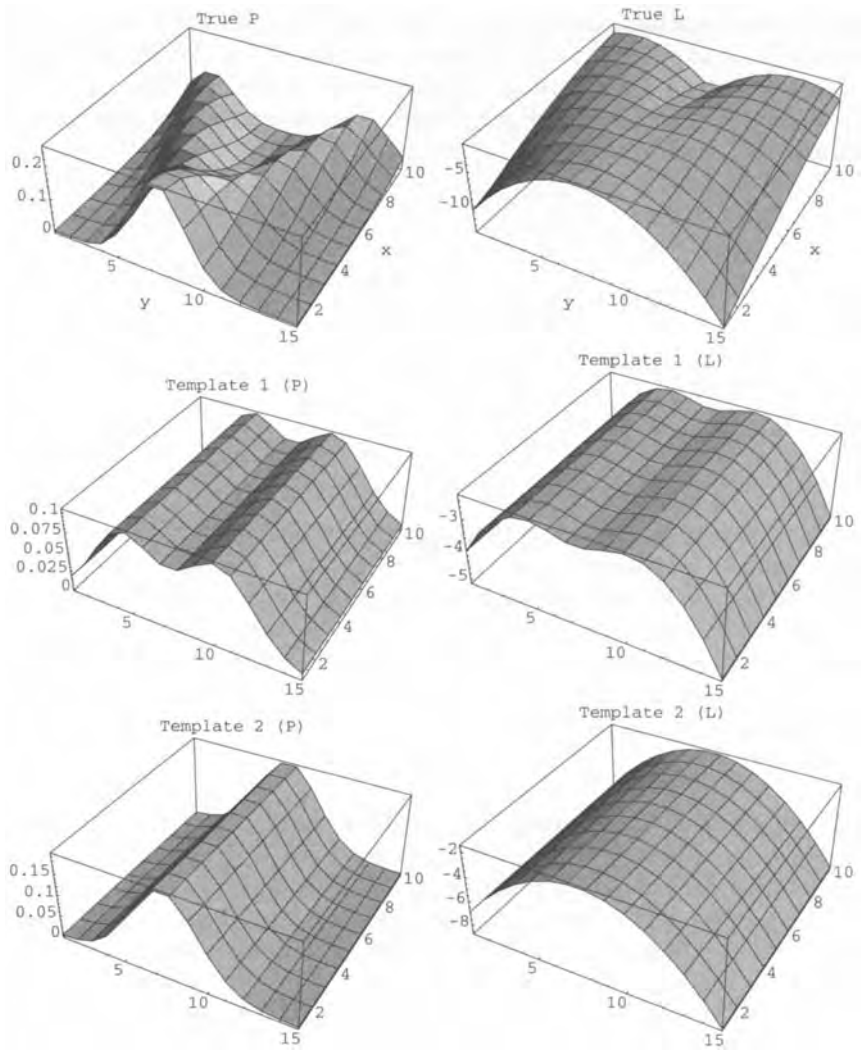


Figure 6.1: Density estimation 7. The true density and templates used for examples. First row: True density P_{true} (left hand side) and true log–density $L_{\text{true}} = \ln P_{\text{true}}$ (right hand side) used for Figs. 6.2 – 6.5 and Fig. 3.5. Second and third row: Two corresponding reference densities (templates) t_1 and t_2 used in Section 6.3 (Figs. 6.2 – 6.5) for P (t_i^P, left hand side) or for L (t_i^L, right hand side), respectively, with $t_i^L = \ln t_i^P$. As reference we give some values of the expected test error $\int dy\, dx\, p(x)p(y|x,\phi_{\text{true}}) \ln p(y|x,\phi)$ under the true $p(y|x,\phi_{\text{true}})$ for uniform $p(x)$ which takes its minimal value of 2.23 for ϕ_{true}, and is 2.56 for template t_1, 2.90 for template t_2, and 2.68 for a uniform P.

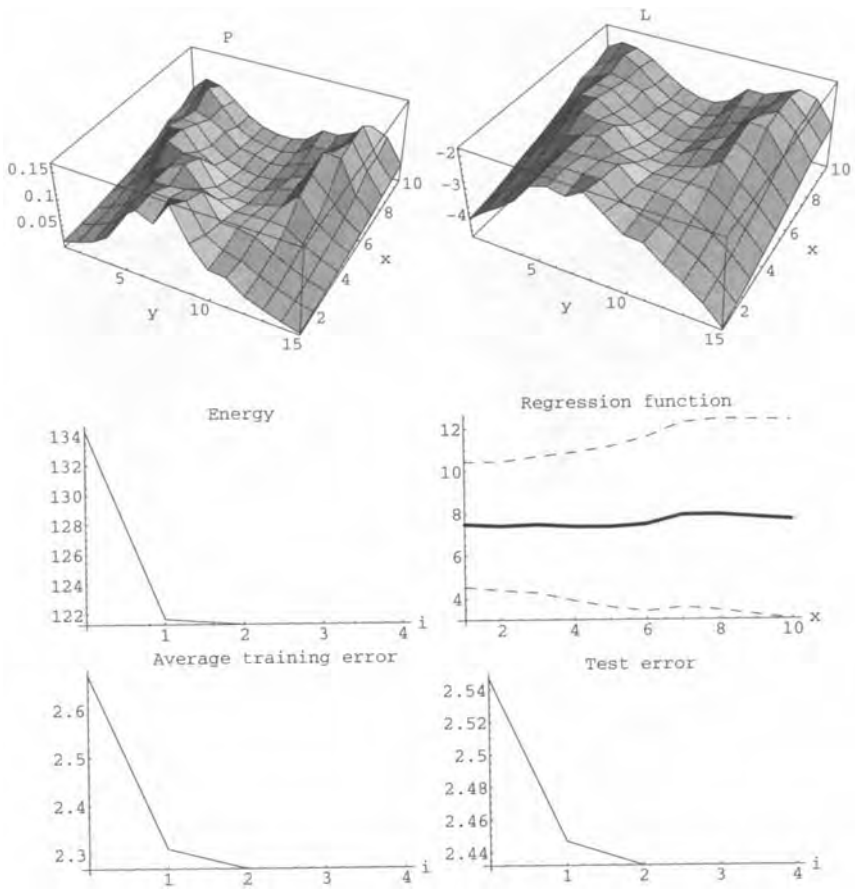


Figure 6.2: Density estimation 8. Density estimation (50 data points) with a Gaussian mixture prior for log–probability L, Laplacian prior, and the two template functions shown in Fig. 6.1. Top row: Final solution $P(x,y) = p(y|x,\phi)$ and $L = \ln P$. Second row: Energy $E(L)$ of (6.10) during iteration and final regression function. Bottom row: Average training error $-(1/n)\sum_{i=1}^{n} \ln p(y_i|x_i,\phi)$ during iteration and average test error $-\int dy\,dx\,p(x)p(y|x,\phi_{\text{true}}) \ln p(y|x,\phi)$ for uniform $p(x)$. (Two mixture components with $\lambda = 0.5$ and smoothness prior with $\mathbf{K}_1 = \mathbf{K}_2$ of the form (3.83) with $\lambda_x = 2$, $\lambda_y = 1$, $\lambda_0 = 0$, $\lambda_2 = 1$, $\lambda_4 = \lambda_6 = 0$, massive prior iteration with $\mathbf{A} = \mathbf{K} + m^2\mathbf{I}$ and squared mass $m^2 = 0.01$, initialized with $L^{(0)} = t_1$. At each iteration step i the factor η has been adapted by a line search algorithm. Mesh with 10 points in x-direction and 15 points in y–direction. Except for the inclusion of two mixture components, parameters are equal to those for Fig. 3.5.)

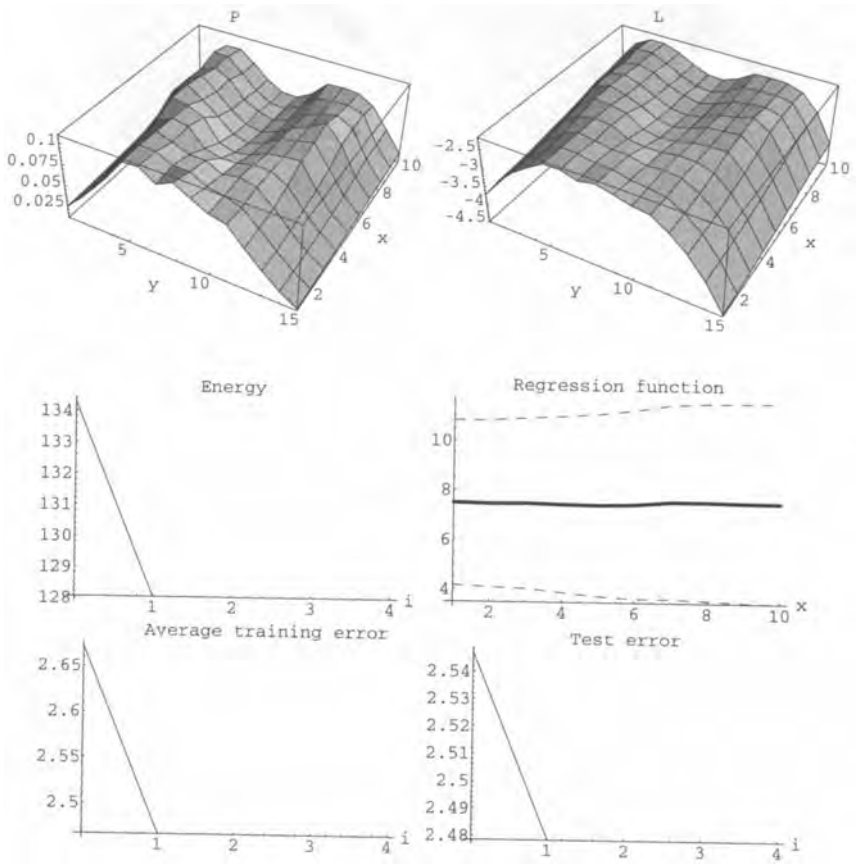


Figure 6.3: Density estimation 9. Density estimation (50 data points) with a Gaussian mixture prior and large λ. Density estimation for log–probability L with the same parameters as for Fig. 6.2, except for $\lambda = 1.0$. Due to the larger smoothness constraint the averaged training error is larger than in Fig. 6.2. The fact that the test error is also somewhat larger than in Fig. 6.2 indicates that the value of λ is already too large. Convergence, however, is very fast.

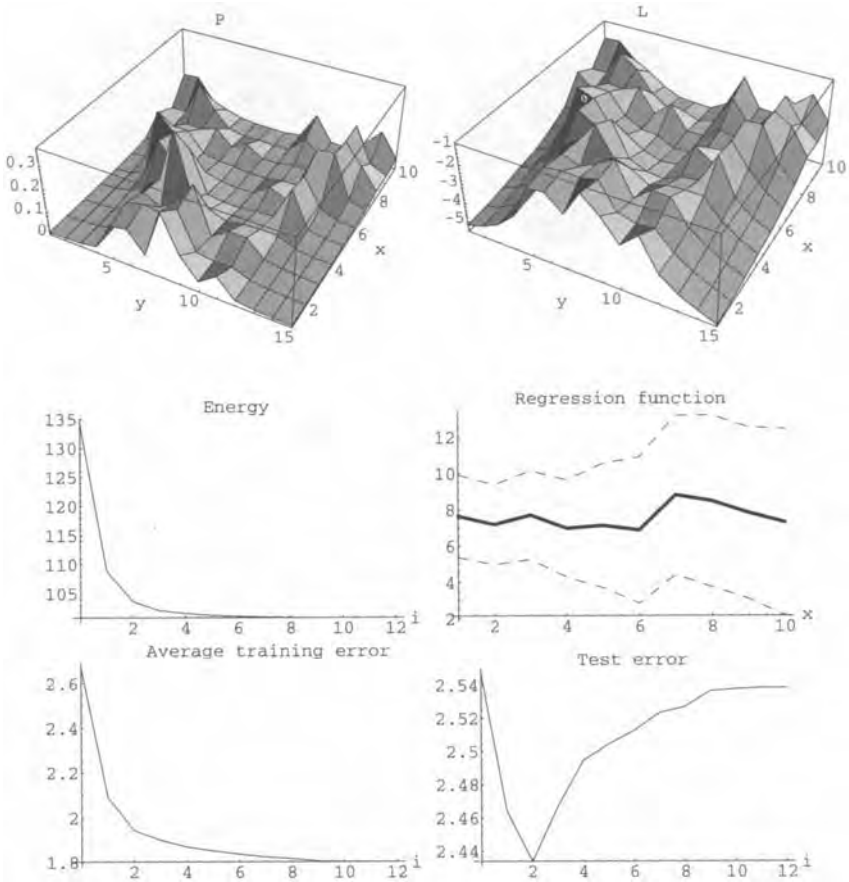


Figure 6.4: Density estimation 10. Density estimation (50 data points) with a Gaussian mixture prior and overfitting due to small λ. (Same parameters as for Fig. 6.2, except for $\lambda = 0.1$.) A small λ allows the average training error to become quite small. However, the average test error grows already after two iterations. (Having found during iteration at some λ–value an increasing test error, often, but not always, a better solution can be found by changing λ.)

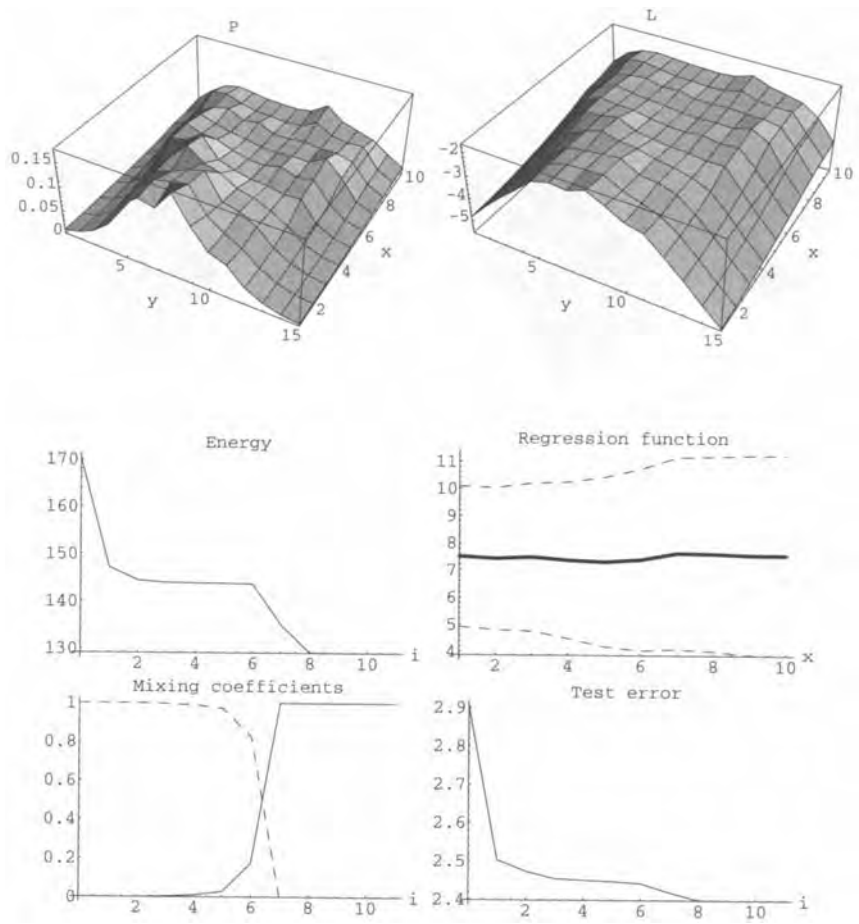


Figure 6.5: Density estimation 11. Gaussian mixture prior: Example of an approximately stable solution. (Same parameters as for Fig. 6.2, except for $\lambda = 1.2$, $m^2 = 0.5$, and initialized with $L = t_2$.) A nearly stable solution is obtained after two iterations, followed by a plateau between iterations 2 and 6. A better solution is finally found with smaller distance to template t_1. The plateau gets elongated with growing mass m. The figure on the left hand side of the bottom row shows the mixing coefficients a_j of the components of the prior mixture model for the solution during iteration (a_1 full line, a_2 dashed line).

6.4 Prior mixtures for regression

6.4.1 General framework

For regression it is useful to introduce an inverse temperature β multiplying the terms depending on ϕ, i.e., likelihood and prior.[1] Similar to Eq. (6.1) we assume

$$p(\phi, \theta|\beta, D, D_0) = \sum_i p(\phi|\theta, j, \beta, D, D_0)\, p(\theta|j, \beta)\, p(j). \tag{6.18}$$

Hence for given β the error or energy functional (6.3) becomes for regression, where ϕ stands for the regression function,

$$E_{\theta,\phi} = -\ln \sum_j^m e^{-\beta E_{\phi,j} - E_{\theta,\beta,j} + c_j} = -\ln \sum_j^m e^{-E_j + c_j}, \tag{6.19}$$

where

$$E_{\phi,j} = E_D + E_{0,j}, \tag{6.20}$$

$$E_j = \beta E_{\phi,j} + E_{\theta,\beta,j} = \beta E_D + \beta E_{0,j} + E_{\theta,\beta,j}, \tag{6.21}$$

$$E_D = \frac{1}{2} \langle \phi - t_D \,|\mathbf{K}_D|\, \phi - t_D \rangle, \tag{6.22}$$

$$E_{0,j} = \frac{1}{2} \langle \phi - t_j(\theta) \,|\mathbf{K}_j(\theta)|\, \phi - t_j(\theta) \rangle, \tag{6.23}$$

with some hyperprior energy $E_{\theta,\beta,j}$ (with arbitrary β–dependency), and also ϕ–independent

$$\begin{aligned} c_j(\theta, \beta) &= -\ln Z_\phi(\theta, j, \beta) + \frac{n}{2} \ln \frac{\beta}{2\pi} - \frac{\beta}{2} V_D - c \\ &= \frac{1}{2} \ln \det \left(\mathbf{K}_j(\theta)\right) + \frac{d+n}{2} \ln \frac{\beta}{2\pi} - \frac{\beta}{2} V_D - c \end{aligned} \tag{6.24}$$

for n data points, a d–dimensional discretization of the ϕ–integral, and some constant c. If we also maximize with respect to β we have to include the (ϕ–independent) training data variance for data at the same x, $V_D = \sum_i^n V_i$ where $V_i = \sum_{k=1}^{n_i} y_k(x_i)^2/n_i - t_D^2(x_i)$ is the variance of the n_i training data at x_i. Clearly, in case every x_i appears only once in the data the variance V_D vanishes. Notice that c_j includes a contribution from the n data points arising from the β–dependent normalization of the likelihood term. Writing the stationarity equation for the hyperparameter β separately, the corresponding

[1] The variable β is, together with the regression function $\phi(x)$, part of the hidden variables $\tilde{\phi} = (\phi, \beta)$ which determine the likelihood, and thus is not a hyperparameter. Technically, however, β can be treated like a regularization factor. For that reason it is included in this chapter on hyperparameters.

three stationarity conditions are found as

$$0 = \sum_j^m a_j \left(\mathbf{K}_D (\phi - t_D) + \mathbf{K}_j (\phi - t_j) \right), \tag{6.25}$$

$$0 = \sum_j^m a_j \left(\beta \frac{\partial E_{0,j}}{\partial \theta} + \frac{\partial E_{\theta,\beta,j}}{\partial \theta} + \frac{1}{2} \mathrm{Tr} \left(\mathbf{K}_j^{-1} \frac{\partial \mathbf{K}_j}{\partial \theta} \right) \right), \tag{6.26}$$

$$0 = \sum_j^m a_j \left(E_{\phi,j} + \frac{\partial E_{\theta,\beta,j}}{\partial \beta} + \frac{d+n}{2\beta} - \frac{1}{2} V_D \right), \tag{6.27}$$

with normalized

$$\begin{aligned} a_j &= p(j|\phi,\theta,\beta,D_0) = \frac{e^{-E_j+c_j}}{\sum_k^m e^{-E_k+c_k}} = \frac{e^{-\beta E_{0,j} - E_{\theta,\beta,j} + \frac{1}{2}\ln\det \mathbf{K}_j}}{\sum_k^m e^{-\beta E_{0,k} - E_{\theta,\beta,k} + \frac{1}{2}\ln\det \mathbf{K}_k}} \\ &= \frac{p(\phi|j,\theta,\beta,D_0) p(j|\theta,\beta,D_0)}{p(\phi|\theta,\beta,D_0)} = \frac{p(\phi|j,\theta,\beta,D_0) p(j,\theta|\beta,D_0)}{p(\phi,\theta|\beta,D_0)}. \end{aligned} \tag{6.28}$$

As β is only a one–dimensional parameter and its density can be quite non–Gaussian it is probably in general more informative to solve for several values of β than to restrict to a single 'optimal' β^*. Eq. (6.25) can also be written

$$\phi = \left(\mathbf{K}_D + \sum_j^m a_j \mathbf{K}_j \right)^{-1} \left(\mathbf{K}_D t_D + \sum_l^m a_j \mathbf{K}_j t_j \right), \tag{6.29}$$

which, due to the ϕ-dependence of a_j, is a nonlinear equation for ϕ.

6.4.2 High and low temperature limits

The limits of large and of small β make the introduction of this additional parameter useful: The high temperature limit $\beta \to 0$ gives the convex case, and statistical mechanics provides us with high and low temperature expansions. Hence we study the high temperature and low temperature limits of Eq. (6.29).

In the *high temperature limit* $\beta \to 0$ the exponential factors a_j become ϕ–independent

$$a_j \overset{\beta \to 0}{\longrightarrow} a_j^0 = \frac{e^{-E_{\theta,\beta,j} + \frac{1}{2}\ln\det \mathbf{K}_j}}{\sum_k^m e^{-E_{\theta,\beta,k} + \frac{1}{2}\ln\det \mathbf{K}_k}}. \tag{6.30}$$

In the special case $E_{\theta,\beta,j} = E_{\beta,j} + \beta E_\theta$, one has to replace $E_{\theta,\beta,j}$ by $E_{\beta,j}$. The high temperature solution becomes

$$\phi = \bar{t}, \tag{6.31}$$

with (generalized) 'complete template average'

$$\bar{t} = \left(\mathbf{K}_D + \sum_j^m a_j^0 \mathbf{K}_j \right)^{-1} \left(\mathbf{K}_D t_D + \sum_l^m a_j^0 \mathbf{K}_j t_j \right). \tag{6.32}$$

Notice that $\bar{t}$ corresponds to the minimum of the quadratic functional

$$E_{(\beta \to 0)} = \frac{1}{2} \langle \phi - t_D \,|\mathbf{K}_D|\, \phi - t_D \rangle + \frac{1}{2} \sum_j^m a_j^0 \langle \phi - t_j \,|\mathbf{K}_j|\, \phi - t_j \rangle . \tag{6.33}$$

Thus, in the infinite temperature limit, a combination of quadratic priors by OR is effectively replaced by a combination by AND.

In the *low temperature limit* $\beta \to \infty$ we have, assuming for the sake of simplicity an additive structure $E_{\theta,\beta,j} = E_\beta + E_{(j)} + \beta E_\theta$, with, e.g., $p(j) \propto \exp(-E_{(j)})$

$$\sum_j e^{-\beta(E_{0,j}+E_\theta)-E_\beta-E_{(j)}} = e^{-\beta(E_{0,j^*}+E_\theta)-E_\beta} \sum_j e^{-\beta(E_{0,j}-E_{0,j^*})-E_{(j)}} \tag{6.34}$$

$$\overset{\beta\to\infty}{\longrightarrow} e^{-\beta(E_{0,j^*}+E_\theta)-E_\beta-E_{(j)}} \quad \text{if} \quad E_{0,j^*} < E_{0,j} \text{ for all } j \neq j^* \text{ and } p(j^*) \neq 0, \tag{6.35}$$

meaning that only the component j^* with minimal energy survives (as $\mathbf{K}_j$ is β–independent)

$$a_j \overset{\beta\to\infty}{\longrightarrow} a_j^\infty = \begin{cases} 1 & : \quad j = j^* = \operatorname{argmin}_j E_{0,j} = \operatorname{argmin}_j E_{\phi,j} \\ 0 & : \quad j \neq \operatorname{argmin}_j E_{0,j} = \operatorname{argmin}_j E_{\phi,j} \end{cases} . \tag{6.36}$$

Henceforth, all (generalized) 'component averages' $\bar{t}_j$ become solutions

$$\phi = \bar{t}_j, \tag{6.37}$$

with

$$\bar{t}_j = (\mathbf{K}_D + \mathbf{K}_j)^{-1} (\mathbf{K}_D t_D + \mathbf{K}_j t_j) \tag{6.38}$$

provided the $\bar{t}_j$ fulfill the stability condition

$$E_{\phi,j}(\phi = \bar{t}_j) < E_{\phi,j'}(\phi = \bar{t}_j), \quad \forall j' \neq j, \tag{6.39}$$

i.e.,

$$V_j < \frac{1}{2} \langle \bar{t}_j - \bar{t}_{j'} \,|\mathbf{K}_D + \mathbf{K}_{j'}|\, \bar{t}_j - \bar{t}_{j'} \rangle + V_{j'}, \quad \forall j' \neq j, \tag{6.40}$$

where

$$V_j = \frac{1}{2} \Big(\langle t_D \,|\mathbf{K}_D|\, t_D \rangle + \langle t_j \,|\mathbf{K}_j|\, t_j \rangle - \langle \bar{t}_j \,|\mathbf{K}_D + \mathbf{K}_j|\, \bar{t}_j \rangle \Big). \tag{6.41}$$

That means single components become solutions at zero temperature $1/\beta$ in case their (generalized) 'template variance' V_j, measuring the discrepancy between data and prior term, is not too large. Eq. (6.29) for ϕ can also be expressed by the component solutions $\bar{t}_j$ (or low temperature solutions if Eq. (6.39) is fulfilled)

$$\phi = \left(\sum_j^m a_j (\mathbf{K}_D + \mathbf{K}_j) \right)^{-1} \sum_j^m a_j \left(\mathbf{K}_D + \mathbf{K}_j \right) \bar{t}_j . \tag{6.42}$$

Summarizing, in the high temperature limit the stationarity equation (6.25) becomes linear with a single solution being a (generalized) average of all template functions. In the low temperature limit the single component solutions become stable provided their (generalized) variance, corresponding to their minimal error, is small enough.

6.4.3 Equal covariances and the ferromagnet

Especially interesting is the case of j–independent $\mathbf{K}_j(\theta) = \mathbf{K}_0(\theta)$ and θ–independent $\det \mathbf{K}_0(\theta)$. In that case the determinants of $\mathbf{K}_j$, often difficult to obtain, do not have to be calculated.

For j–independent inverse covariances the high temperature solution is, according to Eqs. (6.32) and (6.38), a linear combination of the component solutions (or low temperature solutions if Eq. (6.39) is fulfilled),

$$\bar{t} = \sum_j^m a_j^0 \bar{t}_j . \tag{6.43}$$

It is worth emphasizing that, as the solution $\bar{t}$ is *not* a mixture of the component templates t_j but of component solutions $\bar{t}_j$, even poor choices for the template functions t_j can lead to good solutions, if enough data are available.

Eq. (6.42) simplifies to

$$\phi = \frac{\sum_j^m \bar{t}_j e^{-\beta E_{\phi,j}(\phi) - E_{\theta,\beta,j} + c_j}}{\sum_j^m e^{-\beta E_{\phi,j}(\phi) - E_{\theta,\beta,j} + c_j}} = \sum_j^m a_j \bar{t}_j = \bar{t} + \sum_j^m (a_j - a_j^0) \bar{t}_j , \tag{6.44}$$

with, according to Eq. (6.38),

$$\bar{t}_j = \left(\mathbf{K}_D + \mathbf{K}_0 \right)^{-1} \left(\mathbf{K}_D t_D + \mathbf{K}_0 t_j \right) , \tag{6.45}$$

and, for j–independent c_j in Eq. (6.28),

$$a_j = \frac{e^{-E_j}}{\sum_k e^{-E_k}} = \frac{e^{-\beta E_{\phi,j} - E_{\theta,\beta,j}}}{\sum_k e^{-\beta E_{\phi,k} - E_{\theta,\beta,k}}} = \frac{e^{-\frac{\beta}{2} a B_j a + d_j}}{\sum_k e^{-\frac{\beta}{2} a B_k a + d_k}} , \tag{6.46}$$

introducing vector a with components a_j, $(m \times m)$–matrices

$$B_j(k,l) = \langle \bar{t}_k - \bar{t}_j \,|\mathbf{K}_D + \mathbf{K}_0|\, \bar{t}_l - \bar{t}_j \rangle \tag{6.47}$$

and constants

$$d_j = -\beta V_j - E_{\theta,\beta,j}, \tag{6.48}$$

with V_j given in (6.41). Eq. (6.44) is still a nonlinear equation for ϕ, it shows, however, that the solutions must be convex combinations of the ϕ–independent $\bar{t}_j$. Thus it is sufficient to solve Eq. (6.46) for m mixture coefficients a_j instead of the stationarity Eq. (6.25) for the function ϕ.

The high temperature relation Eq. (6.30) becomes

$$a_j \overset{\beta\to 0}{\longrightarrow} a_j^0 = \frac{e^{-E_{\theta,\beta,j}}}{\sum_k^m e^{-E_{\theta,\beta,k}}}, \tag{6.49}$$

or $a_j^0 = 1/m$ for a hyperprior $p(\theta,\beta,j)$ uniform with respect to j. The low temperature relation Eq. (6.36) remains unchanged.

For $m = 2$, Eq. (6.44) becomes

$$\phi = \sum_j^2 a_j \bar{t}_j = \frac{\bar{t}_1 + \bar{t}_2}{2} + (a_1 - a_2)\frac{\bar{t}_1 - \bar{t}_2}{2} = \frac{\bar{t}_1 + \bar{t}_2}{2} + (\tanh \Delta)\,\frac{\bar{t}_1 - \bar{t}_2}{2}, \tag{6.50}$$

with $a_1 + a_2 = 1$ and

$$\begin{aligned} \Delta &= \frac{E_2 - E_1}{2} = \beta \frac{E_{\phi,2} - E_{\phi,1}}{2} + \frac{E_{\theta,\beta,2} - E_{\theta,\beta,1}}{2} \\ &= -\frac{\beta}{4} a(B_1 - B_2)a + \frac{d_1 - d_2}{2} = \frac{\beta}{4} b(2a_1 - 1) + \frac{d_1 - d_2}{2}, \end{aligned} \tag{6.51}$$

because the matrices B_j are in this case zero except $B_1(2,2) = B_2(1,1) = b$. In case $E_{\theta,\beta,j}$ is uniform in j, so that $a_j^0 = 0.5$, it follows $(\bar{t}_1 + \bar{t}_2)/2 = \bar{t}$. The stationarity Eq. (6.46) for the vector a of mixture coefficients, being equivalent to the stationarity equation Eq. (6.25) for ϕ, can be solved graphically (see Figs. 6.6, 6.7). The solution is given by the point where $a_1 e^{-\frac{\beta}{2} b a_1^2 + d_2} = (1 - a_1) e^{-\frac{\beta}{2} b (1-a_1)^2 + d_1}$, or alternatively,

$$a_1 = \frac{1}{2}\,(\tanh \Delta + 1)\,. \tag{6.52}$$

This equation is analogous to the celebrated mean field equation of the ferromagnet. The dependency of a one–dimensional ϕ on β is shown in Fig. 6.8, 6.9 for some examples.

We conclude that in the case of equal component covariances, in addition to the linear low–temperature equations, only an $(m-1)$–dimensional, nonlinear equation has to be solved to determine the 'mixing coefficients' $a_1, \cdots, a_{m-1}$.

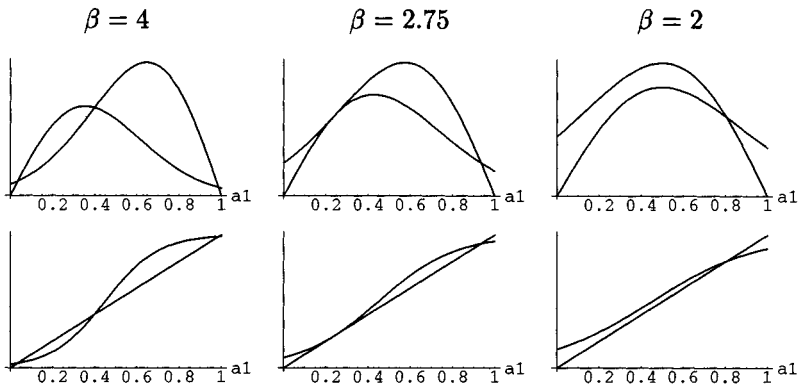


Figure 6.6: Gaussian mixture prior and the ferromagnet. The solution of stationarity equation Eq. (6.46) is given by the point where $a_1 e^{-\frac{\beta}{2}ba_1^2+d_2} = (1-a_1)e^{-\frac{\beta}{2}b(1-a_1)^2+d_1}$ (upper row) or, equivalently, $a_1 = \frac{1}{2}(\tanh\Delta + 1)$ (lower row). Shown are, from left to right, a situation at low temperature ($\beta = 4$) with two stable and one unstable solutions, at a temperature ($\beta = 2.75$) near the bifurcation, and at high temperature ($\beta = 2$) with one stable solution. The values of $b = 2$, $d_1 = -0.2025\beta$ and $d_2 = -0.3025\beta$ used for the plots correspond to the one–dimensional model of Fig. 6.8 with $t_1 = 1$, $t_2 = -1$, $t_D = 0.1$. Notice, however, that the shown relation is valid for $m = 2$ at arbitrary dimension.

6.4.4 Some analytical calculations for mixture models

Like in simple Gaussian regression, the predictive density for regression under a Gaussian mixture model can be calculated analytically for fixed hyperparameter θ. The integral over θ, however, generally has to be done numerically or in MAP.

The predictive density can be expressed in terms of the likelihood of θ and j marginalized (integrated) over ϕ, i.e., $p(y_D|x_D, D_0, \theta, j)$,

$$p(y|x, D, D_0) = \sum_j \int d\theta \, \frac{p(\theta, j)\, p(y_D|x_D, D_0, \theta, j) p(y|x, D, D_0, \theta, j)}{\sum_j \int d\theta \, p(\theta, j) p(y_D|x_D, D_0, \theta, j)}. \quad (6.53)$$

(Here we concentrate on θ, the parameter β can be treated analogously.) Similar to Eq. (5.22) the likelihood can be written

$$p(y_D|x_D, D_0, \theta, j) = e^{-\beta \widetilde{E}_{0,j}(\theta) + \frac{1}{2}\ln\det(\frac{\beta}{2\pi}\widetilde{\mathbf{K}}_j(\theta))}, \quad (6.54)$$

with

$$\widetilde{E}_{0,j}(\theta) = \frac{1}{2}\left\langle t_D - t_j(\theta) \left| \widetilde{\mathbf{K}}_j(\theta) \right| t_D - t_j(\theta) \right\rangle = V_j, \quad (6.55)$$

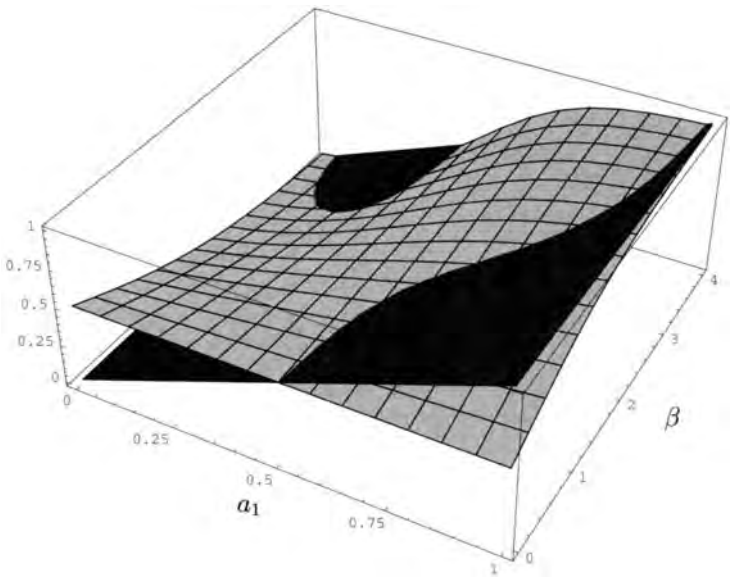


Figure 6.7: Gaussian mixture prior and the ferromagnet. As in Fig. 6.6 the plots of $f_1(a_1) = a_1$ and $f_2(a_1) = \frac{1}{2}(\tanh\Delta + 1)$ are shown within the inverse temperature range $0 \leq \beta \leq 4$.

and

$$\widetilde{\mathbf{K}}_j(\theta) = (\mathbf{K}_D^{-1} + \mathbf{K}_{j,DD}^{-1}(\theta))^{-1}, \tag{6.56}$$

being an $(\tilde{n} \times \tilde{n})$–matrix in data space as in Eq. (3.274). The equality of V_j and $\widetilde{E}_{0,j}$ can be seen using $\mathbf{K}_j - \mathbf{K}_j(\mathbf{K}_D + \mathbf{K}_j)^{-1}\mathbf{K}_j = \mathbf{K}_D - \mathbf{K}_D(\mathbf{K}_D + \mathbf{K}_{j,DD})^{-1}\mathbf{K}_D = \mathbf{K}_{j,DD} - \mathbf{K}_{j,DD}(\mathbf{K}_D + \mathbf{K}_{j,DD})^{-1}\mathbf{K}_{j,DD} = \widetilde{\mathbf{K}}_j$. For the predictive mean, being the optimal solution under squared–error loss and log–loss (restricted to Gaussian densities with fixed variance), we find therefore

$$\bar{y}(x) = \int dy\, y\, p(y|x, D, D_0) = \sum_j \int d\theta\; b_j(\theta)\, \bar{t}_j(\theta), \tag{6.57}$$

with, according to Eq. (3.272),

$$\bar{t}_j(\theta) = t_j + \mathbf{K}_j^{-1}\widetilde{\mathbf{K}}_j(t_D - t_j) \tag{6.58}$$

and mixture coefficients

$$b_j(\theta) \;=\; p(\theta, j|D) = \frac{p(\theta, j)p(y_D|x_D, D_0, \theta, j)}{\sum_j \int d\theta p(\theta, j)p(y_D|x_D, D_0, \theta, j)}$$

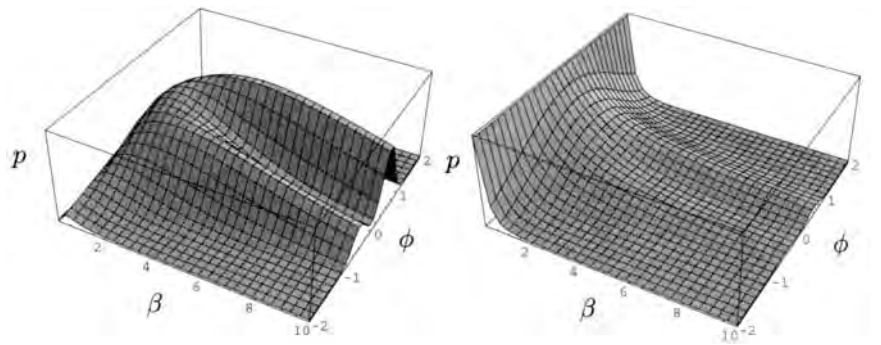


Figure 6.8: Joint posterior density of ϕ and β. Shown is the joint posterior density of ϕ and β given data (D, D_0), $p(\phi, \beta|D, D_0) \propto p(y_D|\phi, \beta)p(\phi|\beta, D_0)p(\beta)$ for an example of a Gaussian prior mixture model with fixed ('zero–dimensional') x, a one–dimensional ϕ–space, training data $y_D = 0.1$, prior data $y_{D_0} = \pm 1$, and inverse temperature β. Left hand side: For uniform hyperprior $p(\beta) \propto 1$ with joint posterior $p \propto e^{-\frac{\beta}{2}\phi^2 + \ln\beta}\left(e^{-\frac{\beta}{2}(\phi-1)^2} + e^{-\frac{\beta}{2}(\phi+1)^2}\right)$ the maximum appears at finite β. (Here no factor $1/2$ appears in front of $\ln\beta$ because normalization constants for prior and likelihood term have to be included.) Right hand side: For compensating hyperprior $p(\beta) \propto 1/\sqrt{\beta}$ with $p \propto e^{-\frac{\beta}{2}\phi^2 - \frac{\beta}{2}(\phi-1)^2} + e^{-\frac{\beta}{2}\phi^2 - \frac{\beta}{2}(\phi+1)^2}$ the maximum is at $\beta = 0$.

$$\propto \quad e^{-\beta\widetilde{E}_{0,j}(\theta) - E_{\theta,j} + \frac{1}{2}\ln\det(\widetilde{K}_j(\theta))}, \tag{6.59}$$

which defines $\widetilde{E}_j = \beta\widetilde{E}_{0,j} + E_{\theta,j}$. For solvable θ–integral the coefficients b_j can therefore be obtained directly from Eq. (6.59).

If b_j is calculated in saddle point approximation at $\theta = \theta^*$ it has the structure of a_j in (6.28) with $E_{0,j}$ replaced by $\widetilde{E}_{0,j}$ and $\mathbf{K}_j$ by $\widetilde{\mathbf{K}}_j$. (The inverse temperature β can be treated analogously to θ. In that case $E_{\theta,j}$ has to be replaced by $E_{\theta,\beta,j}$.)

Calculating also the likelihood for j and θ in Eq. (6.59) in saddle point approximation, where $p(y_D|x_D, D_0, \theta^*, j) \approx p(y_D|x_D, \phi^*)p(\phi^*|D_0, \theta^*, j)$, the terms $p(y_D|x_D, \phi^*)$ in numerator and denominator cancel, so that, skipping D_0 and β,

$$b_j(\theta^*) = \frac{p(\phi^*|j, \theta^*)p(j, \theta^*)}{p(\phi^*, \theta^*)} = a_j(\phi^*, \theta^*) \tag{6.60}$$

becomes equal to the $a_j(\theta^*)$ in Eq. (6.28) at $\phi = \phi^*$.

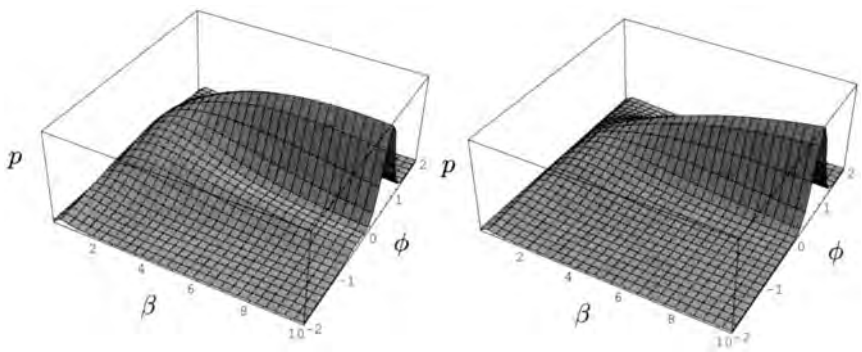


Figure 6.9: Joint posterior density of ϕ and β (different data). Same prior mixture model with fixed x for uniform hyperprior on β as in Fig. 6.8 but for different data $y_D = 0.3$ (left), $y_D = 0.5$ (right).

Eq. (6.59) yields as stationarity equation for θ, similar to Eq. (5.27)

$$0 = \sum_j b_j \left(\frac{\partial \widetilde{E}_j}{\partial \theta} - \frac{1}{2} \mathrm{Tr} \left(\widetilde{\mathbf{K}}_j^{-1} \frac{\partial \widetilde{\mathbf{K}}_j}{\partial \theta} \right) \right) \tag{6.61}$$

$$\begin{aligned} = \sum_j b_j \Bigg[& \beta \left\langle \frac{\partial t_j(\theta)}{\partial \theta} \middle| \widetilde{\mathbf{K}}_j(\theta) \middle| t_j(\theta) - t_D \right\rangle \\ & + \frac{\beta}{2} \left\langle t_D - t_j(\theta) \middle| \frac{\partial \widetilde{\mathbf{K}}_j(\theta)}{\partial \theta} \middle| t_D - t_j(\theta) \right\rangle \\ & - \frac{1}{2} \mathrm{Tr} \left(\widetilde{\mathbf{K}}_j^{-1}(\theta) \frac{\partial \widetilde{\mathbf{K}}_j(\theta)}{\partial \theta} \right) - \frac{1}{p(\theta, j)} \frac{\partial p(\theta, j)}{\partial \theta} \Bigg]. \end{aligned} \tag{6.62}$$

For fixed θ and j–independent covariances the high temperature solution is, according to Eq. (6.43), a mixture of component solutions weighted by their prior

$$\bar{y} \overset{\beta \to 0}{\longrightarrow} \sum_j p(j)\, \bar{t}_j = \sum_j a_j^0\, \bar{t}_j = \bar{t}. \tag{6.63}$$

The low temperature solution becomes the component solution $\bar{t}_j$ with minimal distance between data and prior template

$$\bar{y} \overset{\beta \to \infty}{\longrightarrow} \bar{t}_{j^*} \quad \text{with } j^* = \mathrm{argmin}_j \left\langle t_D - t_j \middle| \widetilde{\mathbf{K}}_j \middle| t_D - t_j \right\rangle. \tag{6.64}$$

Fig. 6.10 compares the exact mixture coefficient b_1 with the dominant solution of the maximum posterior coefficient a_1 for the case of two mixture

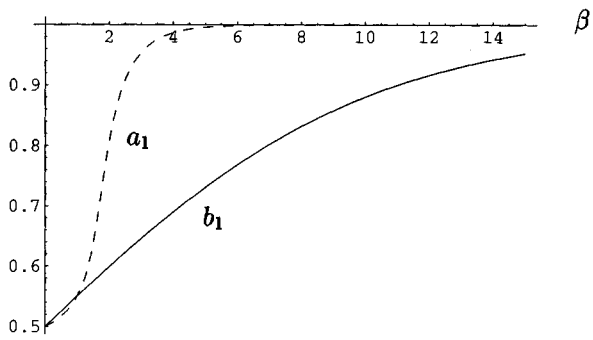


Figure 6.10: Exact and MAP mixture coefficients. Exact b_1 and a_1 (dashed) vs. β for two mixture components with equal covariances and $B_1(2,2) = b = 2$, $\widetilde{E}_1 = 0.405$, $\widetilde{E}_2 = 0.605$.

components with equal covariances discussed in Section 6.4.3 (see also [291]). According to Eq. (6.46) the coefficients a_j and b_j are then related as

$$a_j = \frac{e^{-\frac{\beta}{2}aB_j a - \widetilde{E}_j}}{\sum_k e^{-\frac{\beta}{2}aB_k a - \widetilde{E}_k}} = \frac{b_j\, e^{-\frac{\beta}{2}aB_j a}}{\sum_k b_k\, e^{-\frac{\beta}{2}aB_k a}}. \tag{6.65}$$

6.5 Local mixtures

Global mixture components can be obtained by combining local mixture components. Predicting a time series, for example, one may allow to switch locally (in time) between two or more possible regimes, each corresponding to a different local covariance or template. The problem with combining local alternatives is that the total number of mixture components grows exponentially in the number of local components.

Consider a local prior mixture model, similar to Eq. (5.126),

$$p(\phi|\theta) = e^{-\frac{1}{2}\int dx\, |\omega(x;\theta(x))|^2 - \ln Z_\phi(\theta)}, \tag{6.66}$$

where $\theta(x)$ may be a binary or an integer variable. The local mixture variable $\theta(x)$ labels local alternatives for filtered differences $\omega(x;\theta(x))$ which may differ in their templates $t(x;\theta(x))$ and/or their local filters $\mathbf{W}(x;\theta(x))$. To avoid infinite products, we choose a discretized x–variable (which may include the y–variable for general density estimation problems), so that

$$p(\phi) = \sum_\theta p(\theta) e^{-\frac{1}{2}\sum_x |\omega(x;\theta(x))|^2 - \ln Z_\phi(\theta)}, \tag{6.67}$$

where the sum $\sum_\theta$ is over all local integer variables $\theta(x)$, i.e.,

$$\sum_\theta = \sum_{\theta(x_1)} \sum_{\theta(x_2)} \cdots = \left(\prod_x \sum_{\theta(x)} \right). \tag{6.68}$$

Only for factorizing hyperprior $p(\theta) = \prod_x p\big(\theta(x)\big)$ the complete posterior factorizes

$$\begin{aligned} p(\phi) &= \left(\prod_{x'} \sum_{\theta(x')} \right) \prod_x \left(p\big(\theta(x)\big)\, e^{-\frac{1}{2}|\omega(x;\theta(x))|^2 - \ln Z_\phi(x,\theta(x))} \right) \\ &= \prod_x \sum_{\theta(x)} \left(p\big(\theta(x)\big)\, e^{-\frac{1}{2}|\omega(x;\theta(x))|^2 - \ln Z_\phi(x,\theta(x))} \right), \end{aligned} \tag{6.69}$$

because $\prod_x \sum_{\theta(x)} = \left(\sum_{\theta(x_1)} \sum_{\theta(x_2)} \cdots \right) \prod_x = \sum_\theta \prod_x$ and

$$Z_\phi = \prod_x \sum_{\theta(x)} \left(e^{-\frac{1}{2}|\omega(x;\theta(x))|^2} \right) = \prod_x Z_\phi(x,\theta(x)). \tag{6.70}$$

Under that condition the mixture coefficients a_θ of Eq. (6.28) can be obtained from the equations, local in $\theta(x)$,

$$a_\theta = a_{\theta(x_1)\cdots\theta(x_l)} = p(\theta|\phi) = \prod_x a_{\theta(x)}, \tag{6.71}$$

with

$$a_{\theta(x)} = \frac{p(\theta(x)) e^{-\frac{1}{2}|\omega(x;\theta(x))|^2 - \ln Z_\phi(x;\theta(x))}}{\sum_{\theta'(x)} p(\theta'(x)) e^{-\frac{1}{2}|\omega(x;\theta'(x))|^2 - \ln Z_\phi(x;\theta'(x))}}. \tag{6.72}$$

For equal covariances this is a nonlinear equation within a space of dimension equal to the number of local components. For non–factorizing hyperprior the equations for different $\theta(x)$ can not be decoupled.

6.6 Numerical case study: Image completion

Using mixtures of Gaussian prior processes is especially useful in complex learning situations with a relatively small number of data points where informative but still sufficiently flexible *a priori* information has to be implemented. As a numerical example of such a regression mixture model we study in this section the problem of completing a partially known image. In this context the regression function $\phi(x) = \phi(x_{(1)}, x_{(2)})$ stands for a function of grey levels representing the unknown image to be reconstructed. The discrete, two–dimensional x–variable labels the pixels of the image. The problem

fits into the framework of Gaussian regression with a likelihood of the form (3.164), if we choose as data D grey level values y_i measured with Gaussian noise at pixels $x_i = (x_{(1),i}, x_{(2),i})$. In the following we study such local noisy measurements at single pixels x. It is straightforward to work with nonlocal Gaussian measurements, which, for example, average grey level values over a certain neighborhood of a pixel, by using (3.168) instead of (3.164). Positivity constraints do not have to be implemented explicitly, if the function $\phi(x)$ is mapped before displaying the image into a fixed interval $[0, M]$, with $M > 0$ representing some maximal grey level, by rescaling ϕ and adding or subtracting a constant. This also means that (non–constant) images always use the full range of available grey levels.

If the number of data points is small compared to the total number of pixels, such image reconstruction problems rely essentially on the implementation of 'informative' *a priori* information. In such situations the twofold role of training data becomes especially clear: They do not only represent local information on single grey level values but they provide also, in combination with the prior process, the nonlocal information necessary for generalization. In prior mixture models, for example, the training data determine the *a posteriori* weights of the mixture components, i.e., they select the relevant parts of the prior process.

Now let us consider training data drawn from an original image with Gaussian noise, like those shown in Fig. 6.11. Clearly, a meaningful reconstruction of such a partially known image requires quite specific *a priori* information. Hence, assume we expect the original image to show one of the two faces displayed in Fig. 6.12. Those two reference faces are then the natural choices for the two mean functions $t_j(x)$ of a two–component Gaussian mixture prior. Furthermore, it is not very realistic to represent a reference face $t_j(x)$ by only one fixed image. The same face may appear different on a different image, for example, translated, rotated, or in another distance. This can be modeled by including hyperparameters parameterizing such possible deformations of templates. In particular, we have chosen, for each prior component separately, translations (θ_1, θ_2), scale factors θ_3, θ_4, and a rotation angle (around template center) θ_5, as shown in Fig. 6.13. For a two–component mixture prior this results in a 10-dimensional hyperparameter vector. A more general approach would be to allow optical flow fields parameterizing general smooth local deformations of a reference image as used for warping or morphing an image [172, 233, 524].

Due to the large number of different pixels in the training data ($\tilde{n} \approx 1000$) it was easier to solve Eq. (6.25) by iteration for the full image than to invert the matrix in (6.56) in the $\tilde{n}$–dimensional space of the training data. The prior operator $\mathbf{K}_0$ has been implemented as a 3×3 negative Laplacian filter [173, 272].[2] The regularization factor multiplying the negative Laplacian

[2]Notice the difference to a simple template matching based on a local squared error

Figure 6.11: Image completion 1: Training data. Example of training data ($n = 10000$) for an image completion problem. Shown is the training data template $t_D(x)$, giving for each pixel x the mean of the sampled grey levels y. At pixels for which no data are available $t_D(x)$ has been set to zero, shown as black.

has been set to one. Being a prior dominated task, the solutions do not depend sensitively on that value. Completed images $\phi(x)$ have been obtained by iterating according to

$$\phi^{(k+1)} = \phi^{(k)} + \eta \mathbf{A}^{-1}\Big(\mathbf{K}_D(t_D - \phi^{(k)}) + \mathbf{K}_0\Big(\sum_j a_j^{(k)} t_j - \phi^{(k)}\Big)\Big), \quad (6.73)$$

performed alternating with brute force θ–minimization. A Gaussian matrix $\mathbf{A}^{-1}$ (implemented by a 5×5 binomial filter) proved to be successful (see Appendix C). A good choice for the relaxation factor η has been 0.05, if not adapted during learning.

Reconstructed images have been calculated for different 'inverse mixture temperatures' β. A typical low temperature situation is shown in Fig. 6.14. Like in Fig. 6.7 there exist two local maxima of the posterior, one for each mixture component, with hyperparameters adapted optimally to the given training data. The two solutions are essentially generalized averages (with respect to the metric defined by the negative Laplacian inverse prior covariances of the components) of the training data and of one of the two optimally adapted reference images $t_j(x, \theta^*)$. The influence of the non–dominant reference image is in this case negligible. Fig. 6.15 shows that the situation is

instead of the Laplacian kernel chosen here. The latter results in a smooth interpolation between data and templates. Cf., for example, [66].

different at high temperatures. In that case only one solution exists to which training data and both reference images contribute.

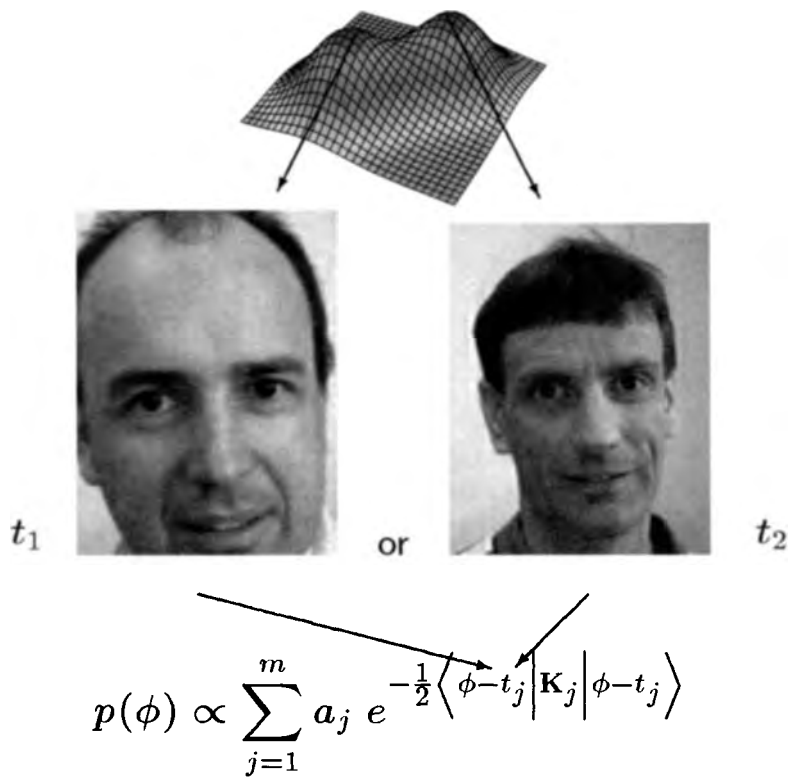


Figure 6.12: Image completion 2: Gaussian mixture prior. Example of a two–component Gaussian mixture prior for image completion. Shown are the mean (reference, template) functions $t_j(x)$ of the two prior components.

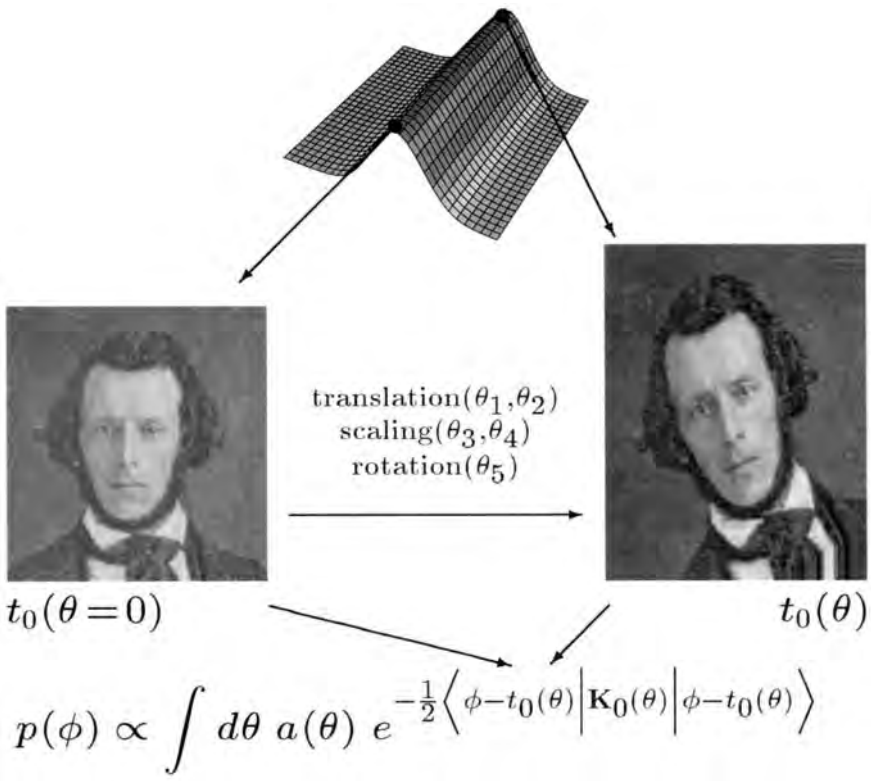


Figure 6.13: Image completion 3: Prior with hyperparameters. Gaussian prior factor for image completion with mean $t_0(x, \theta)$ parameterized by a five dimensional hyperparameter vector θ representing translation, rotation, and rescaling of an image.

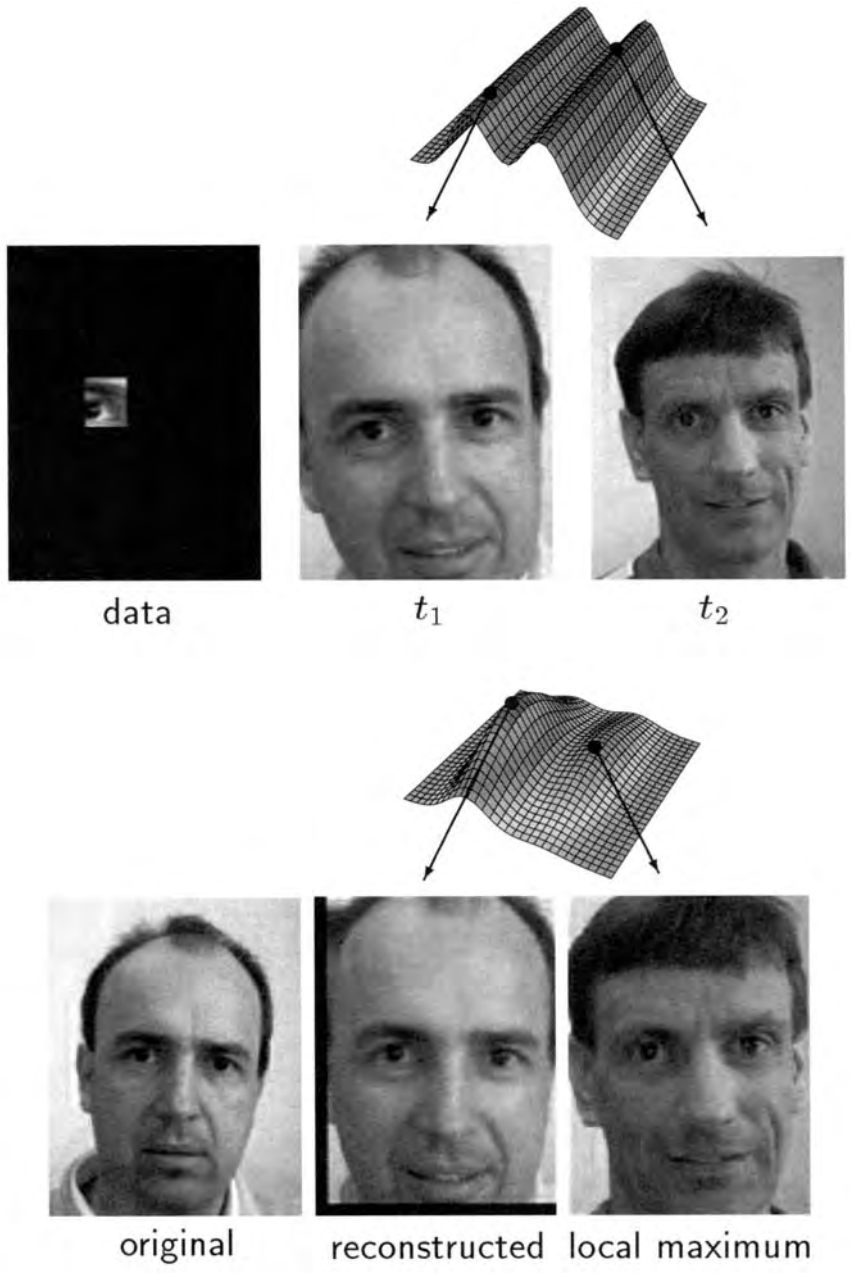


Figure 6.14: Image completion 4: Low temperature solution. Shown are the training data (sampled with Gaussian noise with $\sigma = 10$), the two reference images $t_j(x, \theta)$ (for $\theta = 0$), the original image, the global maximum (labeled 'reconstructed'), and the second local maximum of the posterior at $\beta = 10$ (low temperature), each for optimally adapted hyperparameter vector θ.

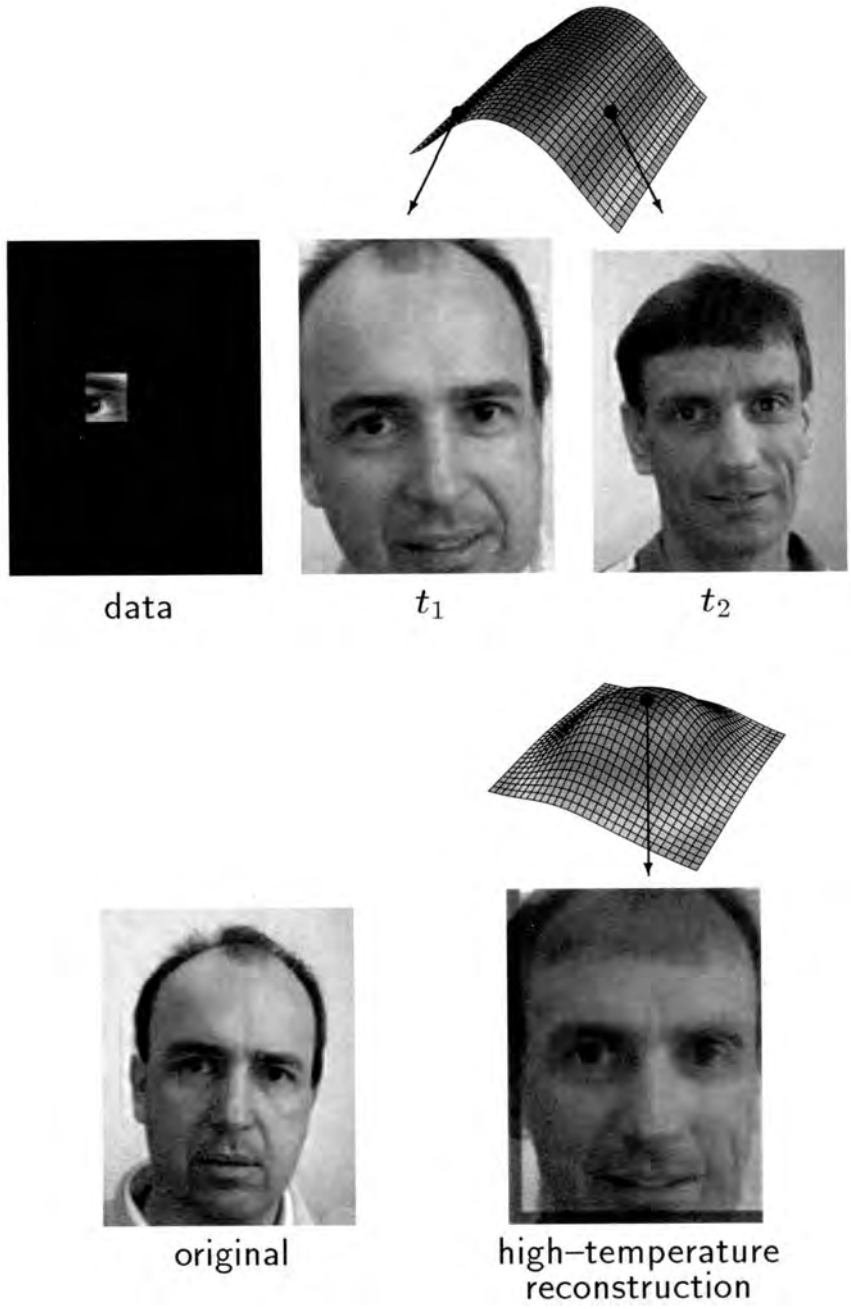


Figure 6.15: Image completion 5: High temperature solution. At high temperature the posterior has only one local (and global) maximum. (Same data and prior as for Fig. 6.14, but with $\beta = 10^{-5}$.)

Chapter 7

Bayesian inverse quantum theory (BIQT)

The problem addressed in this chapter is the reconstruction of the Hamiltonians or potentials of quantum systems from observational data. Finding such 'causes' or 'laws' from a finite number of observations constitutes an *inverse problem* and is typically *ill–posed* in the sense of Hadamard [192,216, 255,262,315,482,483,492]. Two classical research fields dealing with the reconstruction of potentials are *inverse scattering theory* and *inverse spectral theory* [76,77,148,247,262,329,367,465,535]. They characterize the kind of data which are necessary, in addition to a given spectrum, to identify a potential uniquely. For example, such data can be a second complete spectrum for different boundary conditions, knowledge of the potential on a half interval, or the phase shifts as a function of energy. However, neither a complete spectrum nor specific values of potentials or phase shifts for all energies can be determined empirically by a finite number of measurements. Hence any practical algorithm for reconstructing potentials from data must rely on additional *a priori* assumptions, if not explicitly then implicitly. Furthermore, besides energy, other observables like particle coordinates or momenta may have been measured for a quantum system. Therefore, the approach of Bayesian Inverse Quantum Theory (BIQT) developed in the following, is not designed exclusively for spectral data but is able to work with quite arbitrary observational data. In particular, it allows to include situation–specific *a priori* information in a flexible and explicit manner. It can thus easily be adapted to a large variety of different reconstruction scenarios [294,298,304–306,489].

For classical inverse problems Bayesian methods are now in widespread use, like in medicine for X–ray computer tomography [362] or for positron emission tomography, in geophysics to infer the Earth's subsurface properties from reflection seismic data [181,478], and for a variety of other data analysis

problems in astronomy, astrophysics [314], high energy physics [95], and image reconstruction [152,524].

In contrast to the situation for classical inverse problems Bayesian methods for solving inverse quantum problems are not yet well developed. There are, however, good reasons to expect that the development of algorithms for solving inverse problems for quantum systems will become important in the near future. An example is atomic force microscopy when scales are reached where a quantum mechanical description becomes mandatory [155]. Another inverse quantum problem appears in quantum computing when reading off the result of a calculation by measuring the final state of a quantum computer [409,470]. While there already exists some interesting work using the Bayesian framework for the reconstruction of density operators for spin systems [25,71,476], in the following study we will look at the reconstruction of potentials for several different types of quantum systems. We will see that application of the Bayesian framework to the inverse quantum problem of reconstructing potentials is relatively straightforward. Computationally, however, Bayesian methods for inverse quantum problems turn out to be more demanding than, for example, typical applications to regression problems.

In the following we will discuss some examples of BIQT: Bayesian inverse quantum statistics (BIQS) for systems at finite temperature [305], Bayesian inverse time–dependent quantum mechanics (BITDQ) for time–dependent problems [298], and Bayesian inverse Hartree–Fock theory (BIHF) as a mean field approximation for many–body systems [304].

7.1 Bayesian inverse quantum statistics (BIQS)

7.1.1 The likelihood model of quantum theory

We will now study the application of Bayesian methods to inverse problems in quantum statistics [305]. The state of a quantum mechanical system is characterized by its density operator ρ. In particular, the probability (density) of measuring value x for observable O in a state described by ρ is[1] [23,325,

[1]There exist many interpretations of the measurement process in quantum mechanics [20,31]. In the classical interpretation a measurement in quantum mechanics consists of a collapse of the quantum mechanical wave function (projection to an eigenstate of the observable), followed by reading off the measurement result from a classical measurement device [503]. However, the point ('Heisenberg cut'), where to switch from a quantum mechanical description of the system to a classical description, is not well defined. For example, one may model the measurement process as an specific quantum mechanical interaction with a quantum environment. If the interaction is chosen adequately, the states of the environment ('pointer states'), which are measured classically, reflect the states of the measured system. If the 'pointer states' of the environment are (nearly) orthogonal then 'tracing out' the environment yields (nearly) Eq. (7.1). This phenomenon is known as 'decoherence' [164]. In this chapter we use Eq. (7.1), but we do not rely on any specific interpretation of measurement.

	ρ
general pure state	$\|\psi><\psi\|$
stationary pure state	$\|\phi_\alpha(H)><\phi_\alpha(H)\|$
ground state	$\|\phi_0(H)\|><\phi_0(H)\|$
time–dependent pure state	$\|U(t,t_0)\psi(t_0)><U(t,t_0)\psi(t_0)\|$
scattering	$\lim_{\substack{t\to\infty \\ t_0\to-\infty}} \|U(t,t_0)\psi(t_0)><U(t,t_0)\psi(t_0)\|$
general mixture state	$\sum_k p(k)\,\|\psi_k><\psi_k\|$
stationary mixture state	$\sum_i p(i\|H)\,\|\phi_\alpha(H)><\phi_\alpha(H)\|$
canonical ensemble	$(\mathrm{Tr}\, e^{-\beta H})^{-1} e^{-\beta H}$

Table 7.1: Examples of density operators for quantum systems. In this table ψ denotes an arbitrary wave function (representing a pure state), ϕ_α represents an eigenstate of Hamiltonian H with energy E_α. The unitary time evolution operator for a time–independent Hamiltonian H is given by $U = e^{-i(t-t_0)H}$. In scattering one imposes additional specific boundary conditions on the initial and final states [366, 479]. The general mixture state represents the most general type of density operator.

336, 503]

$$p(x|O,\rho) = \mathrm{Tr}\Big(P_O(x)\,\rho\Big). \tag{7.1}$$

This defines the likelihood model of quantum theory. The observable O, represented by a hermitian operator, corresponds to the independent variable x of the previous sections. The projector $P_O(x) = \sum_l |x_l\rangle\langle x_l|$ projects on the space of eigenfunctions $|x_l\rangle$ of O with eigenvalue x, i.e., for which $O|x_l\rangle = x|x_l\rangle$. For non–degenerate eigenvalues $P_O(x) = |x\rangle\langle x|$. Table 7.1 shows some examples of commonly used density operators, which can all be used in a likelihood model (7.1) for a BIQT.

To be specific, we will consider the measurement of particle positions, where $O = \hat{x}$ is the multiplication operator in coordinate space. However, the formalism we will develop does not depend on the particular kind of measured observable. It would be possible, for example, to mix measurements of position

and momentum (for energy measurements see Section 7.1.2).

For the sake of simplicity, we will assume that no *classical noise* is added by the measurement process. It is straightforward, however, to include a classical noise factor in the likelihood function. If, for example, the classical noise is, conditioned on x_i, independent of the quantum system then

$$p(\bar{x}_i|O,\rho) = \int dx_i\, p(\bar{x}_i|x_i)\, p(x_i|O,\rho), \tag{7.2}$$

where we denoted the 'true' coordinates by x_i and the corresponding noisy outcome by $\bar{x}_i$. A simple model for $p(\bar{x}_i|x_i)$ could be a Gaussian.[2]

In contrast to the (ideal) measurement of a classical system, the measurement of a quantum system changes the state of the system. In particular, the measurement process acts as projection $P_O(x)$ on the space of eigenfunctions of operator O with eigenvalues consistent with the measurement result. Thus performing multiple measurements under the assumption of a constant density operator ρ requires special care to ensure the correct preparation of the quantum system before each measurement. Considering a quantum statistical system at finite temperature, as we will study in the next section, the time between two consecutive measurements should be large enough to allow thermalization of the system.

Let us consider a quantum mechanical canonical ensemble at temperature $1/\beta$. Such a system is described by a density operator

$$\rho = \frac{e^{-\beta H}}{\mathrm{Tr}\, e^{-\beta H}}, \tag{7.3}$$

H denoting the Hamiltonian of the system. Specifically, we will focus on repeated measurements of a single particle in a heat bath of temperature $1/\beta$ with sufficient time between measurements to allow the heat bath to reestablish the canonical density operator. For (non–degenerate) particle coordinates x_i the likelihood for ρ becomes the thermal expectation

$$p(x_i|\hat{x},\rho) = \mathrm{Tr}\,(P_{\hat{x}}(x_i)\rho) = \sum_\alpha p_\alpha |\phi_\alpha(x_i)|^2 = < |\phi(x_i)|^2 >, \tag{7.4}$$

$< \cdots >$ denoting the thermal expectation with probabilities

$$p_\alpha = \frac{e^{-\beta E_\alpha}}{Z}, \quad Z = \sum_\alpha e^{-\beta E_\alpha}, \tag{7.5}$$

[2]In this chapter we adopt the notation commonly used in quantum theory and use x to denote coordinates and ϕ (or ψ) to denote wave functions. Compared to the discussion of general density estimation and regression in the previous sections, in this section the hidden states (denoted by ϕ in the previous sections) are represented by a density operator ρ (or by a potential $V(x,x')$ or $v(x)$ which determines ρ), the operator O plays the role of an independent variable (called x in the previous sections), so the measured values (e.g., coordinates) (called y in the previous sections), can be denoted by x, and the symbol ϕ is free to denote a wave function. In short: $\phi \to \rho, V, v$; $x \to O$; $y \to x$.

and energies and orthonormalized eigenstates

$$H\,|\phi_\alpha\rangle = E_\alpha\,|\phi_\alpha\rangle\,. \tag{7.6}$$

We will consider a hermitian Hamiltonian of the standard form $H = T + V$, with kinetic energy T, being $1/(2m)$ times the negative Laplacian $-\Delta$ for a particle with mass m (setting $\hbar = 1$), and local potential $V(x, x') = v(x)\delta(x - x')$. Thus in one dimension

$$H(x, x') = \left(-\frac{1}{2m}\frac{\partial^2}{\partial x^2} + v(x)\right)\delta(x - x'), \tag{7.7}$$

where the δ–functional is usually skipped.

For n independent position measurements x_i, the likelihood for $\rho(v)$ becomes

$$p(x_T|\hat{x}, v) = \prod_{i=1}^{n} p(x_i|\hat{x}, v) = \prod_{i=1}^{n} < |\phi(x_i)|^2 >, \tag{7.8}$$

replacing $p(x_i|\hat{x}, \rho)$ by $p(x_i|\hat{x}, v)$. We remark that it is straightforward to allow β to vary between measurements.

A maximum likelihood approach selects the potential v with maximal likelihood $p(x_T|\hat{x}, v)$ under the training data. Beginning with a discussion of the *parametric* approach we consider a potential $v(\xi, x)$ parameterized by a parameter vector ξ with components ξ_l. To find the parameter vector which maximizes the training likelihood we have to solve the stationarity equation

$$0 = \partial_\xi p(x_T|\hat{x}, v), \tag{7.9}$$

$\partial_\xi = \partial/\partial\xi$ denoting the gradient operator with components ∂_{ξ_l}. From Eq. (7.8) we obtain

$$\begin{aligned} \partial_\xi p(x_i|\hat{x}, v) &= < \big(\partial_\xi \phi^*(x_i)\big)\phi(x_i) > + < \phi^*(x_i)\big(\partial_\xi \phi(x_i)\big) > \\ &- \beta\left(< |\phi(x_i)|^2 \partial_\xi E > - < |\phi(x_i)|^2 >< \partial_\xi E >\right). \end{aligned} \tag{7.10}$$

Hence to solve Eq. (7.9), we have to calculate the derivatives of the eigenvalues $\partial_\xi E_\alpha$ and of the eigenfunctions at the data points $\partial_\xi \phi_\alpha(x_i)$; those are implicitly defined by the eigenvalue equation for $H = H(v)$. To proceed we take the derivative of the eigenvalue equation (7.6)

$$(\partial_\xi H)\,|\phi_\alpha\rangle + H\,|\partial_\xi \phi_\alpha\rangle = (\partial_\xi E_\alpha)\,|\phi_\alpha\rangle + E_\alpha\,|\partial_\xi \phi_\alpha\rangle\,, \tag{7.11}$$

or

$$(E_\alpha - H)\,|\partial_\xi \phi_\alpha\rangle = (\partial_\xi H - \partial_\xi E_\alpha)\,|\phi_\alpha\rangle\,, \tag{7.12}$$

where $\partial_\xi H = \partial_\xi v$. To obtain $\partial_\xi E_\alpha$ we project Eq. (7.12) onto $\langle\phi_{\alpha'}|$, so that for $E_{\alpha'} = E_\alpha$ the hermitian conjugate of Eq. (7.6) yields

$$0 = \langle\phi_{\alpha'}\,|\partial_\xi H - \partial_\xi E_\alpha|\,\phi_\alpha\rangle = \langle\phi_{\alpha'}\,|\partial_\xi H|\,\phi_\alpha\rangle - \partial_\xi E_\alpha\,\langle\phi_{\alpha'}|\,\phi_\alpha\rangle\,, \tag{7.13}$$

from which follows

$$\partial_\xi E_\alpha = \frac{\langle \phi_\alpha | \partial_\xi H | \phi_\alpha \rangle}{\langle \phi_\alpha | \phi_\alpha \rangle}. \tag{7.14}$$

First, we notice that Eq. (7.12) does not determine $\partial_\xi \phi_\alpha$ uniquely because all orbitals with energy E_α (which may be more than one if E_α is degenerate) are in the null space of the operator $(E_\alpha - H)$. Second, Eq. (7.13) shows that the right hand side of Eq. (7.12) possesses no components in the null space of $(E_\alpha - H)$; it represents therefore the solvability condition (3.237) for Eq. (7.12). Third, it is easy to verify that

$$G_\alpha = \sum_{\gamma \neq \alpha} \frac{|\phi_\gamma\rangle \langle \phi_\gamma|}{E_\alpha - E_\gamma}, \tag{7.15}$$

is the Moore–Penrose inverse of $(E_\alpha - H)$, fulfilling $G_\alpha(E_\alpha - H)G_\alpha = G_\alpha$, as discussed in Section 3.8.1. Fourth, a unique solution for $\partial_\xi \phi_\alpha$ can be obtained by fixing its component in the null space of $(E_\alpha - H)$, e.g., by setting $\langle \phi_{\alpha'} | \partial_\xi \phi_\alpha \rangle = 0$ for eigenfunctions $\phi_{\alpha'}$ with $E_{\alpha'} = E_\alpha$. This corresponds to fixing normalization and phase of eigenfunctions and, in case of degenerate eigenvalues, uses the freedom to work with arbitrary, orthonormal linear combinations of the corresponding eigenfunctions. Because the operator $(E_\alpha - H)$ is invertible in the space spanned by all eigenfunctions ϕ_γ with different energy $E_\gamma \neq E_\alpha$, we thus find for orthonormal eigenfunctions,

$$|\partial_\xi \phi_\alpha\rangle = \sum_{\substack{\gamma \\ E_\gamma \neq E_\alpha}} \frac{1}{E_\alpha - E_\gamma} |\phi_\gamma\rangle \langle \phi_\gamma | \partial_\xi H | \phi_\alpha \rangle . \tag{7.16}$$

For non–degenerate energies the sum becomes $\sum_{\gamma \neq \alpha}$. The stationarity equation (7.9) can now be solved iteratively by starting from an initial guess v^0 for v, calculating $E_\alpha(v)$ and $\phi_\alpha(v)$ to obtain $\partial_\xi E_\alpha$ and $\partial_\xi \phi_\alpha(x_i)$ from Eqs. (7.14) and (7.16) and thus $\partial_\xi p(x_i|\hat{x}, v)$ from Eq. (7.10). Then a new guess for v is calculated, switching to log–likelihoods,

$$v^{\text{new}} = v^{\text{old}} + \eta A^{-1} \sum_i \partial_\xi \ln p(x_i|\hat{x}, v^{\text{old}}), \tag{7.17}$$

with some step width η and some positive definite operator A approximating, for example, the Hessian of $\ln p(x_T|\hat{x}, v)$. This procedure is iterated till convergence.

While a parametric approach restricts the space of possible potentials v, a *nonparametric* approach treats each function value $v(x)$ itself as an individual degree of freedom, not restricting the space of potentials. The corresponding nonparametric BIQS stationarity equation is obtained analogous to the parametric stationarity equation (7.9), replacing partial derivatives ∂_ξ by the functional derivative operator $\delta_v = \delta/\delta v$ with components $\delta_{v(x)} = \delta/(\delta v(x))$ [80].

The functional derivative of H with respect to $v(x)$ is now simply

$$\delta_{v(x)} H(x', x'') = \delta_{v(x)} V(x', x'') = \delta(x - x')\delta(x' - x''). \tag{7.18}$$

Using the Moore–Penrose inverse G_α of Eq. (7.15) we get,

$$\delta_{v(x)} \phi_\alpha(x') = \sum_{\substack{\gamma \\ E_\gamma \neq E_\alpha}} \frac{1}{E_\alpha - E_\gamma} \phi_\gamma(x') \phi_\gamma^*(x) \phi_\alpha(x), \tag{7.19}$$

$$\delta_{v(x)} E_\alpha = \frac{\langle \phi_\alpha | \delta_{v(x)} H | \phi_\alpha \rangle}{\langle \phi_\alpha | \phi_\alpha \rangle} = |\phi_\alpha(x)|^2. \tag{7.20}$$

Like for Eq. (7.16) we have set the arbitrary contributions from the null space of $E_\alpha - H$ to zero, $\langle \delta\phi_\alpha | \phi_\alpha \rangle = 0$. This corresponds for real $\phi_\alpha(x)$ to fixing the norm by requiring $\delta_v \langle \phi_\alpha | \phi_\alpha \rangle = 0$, while for complex $\phi_\alpha(x)$ the phases are also controlled.

For the functional derivative of the BIQS–likelihood we get therefore

$$\begin{aligned} \delta_{v(x)} p(x_i|\hat{x}, v) &= < \left(\delta_{v(x)} \phi^*(x_i)\right) \phi(x_i) > + < \phi^*(x_i) \delta_{v(x)} \phi(x_i) > \\ &- \beta \left(< |\phi(x_i)|^2 |\phi(x)|^2 > - < |\phi(x_i)|^2 >< |\phi(x)|^2 > \right). \end{aligned} \tag{7.21}$$

(The partial derivative with respect to parameters ξ and the functional derivative with respect to $v(x)$ are related according to the chain rule $\partial_{\xi_l} p(x_i|\hat{x}, v) = \int dx \, (\partial_{\xi_l} v(x)) \, \delta_{v(x)} p(x_i|\hat{x}, v) = \mathbf{v}_\xi \delta_v p(x_i|\hat{x}, v)$, in operator notation, with operator $\mathbf{v}_\xi(l, x) = \partial_{\xi_l} v(\xi, x)$.)

The large flexibility of nonparametric approaches allows an optimal adaption of v to the available training data. However, as is well known in the context of learning, it is the same flexibility which makes a satisfactory generalization to non–training (future) data impossible, leading, for example, to 'pathological', δ-functional like solutions. Nonparametric approaches require therefore additional restrictions in form of *a priori* information. In the next section we will include *a priori* information in the form of stochastic processes, similar to Bayesian statistics with Gaussian processes [111, 297, 309, 363, 364, 507, 521, 522] or to classical regularization theory [482, 492, 507]. *A priori* information will be implemented *explicitly* in terms of the function values $v(x)$. This is a great advantage over parametric methods where *a priori* information is *implicit* in the chosen parameterization, thus typically difficult or impossible to analyze and not easily adapted to the situation under study. Indeed, because it is only *a priori* knowledge which relates training to non–training data, its explicit control is essential for any successful learning.

7.1.2 Prior models for potentials

Typical results of inverse spectral theory show that a one–dimensional local potential can be reconstructed if, for example, a set of two complete spectra

$\{E_\alpha^{(1)}\}$, $\{E_\alpha^{(2)}\}$ is given for two different boundary conditions for eigenstates ϕ_α [76,262]. Alternatively, a single spectrum is sufficient, if either a complete set of norming constants $u_\alpha = \int \phi_\alpha dx$ is given (for a certain normalization of ϕ_α which fixes the values of ϕ_α on the boundary) [148] or the potential is already known on half of the interval [214]. Results from inverse scattering theory, for example, show under which circumstances a potential can be reconstructed from a complete spectrum and the phase shifts as function of energy [76, 77,367]. In practice, however, the number of actual measurements is finite. Thus even if noiseless measurement devices would be available, an empirical determination of a complete spectrum, or of phase shifts as function of energy, is impossible. Therefore, to reconstruct a potential from experimental data in practice, inverse spectral or inverse scattering theory has to be combined with additional *a priori* information. If such *a priori* information is not made explicit, it nevertheless enters any algorithm at least implicitly.

We address in this chapter the measurement of arbitrary quantum mechanical observables, not restricted to spectral or scattering data. In particular, we consider the measurement of particle positions. However, measuring particle positions only does not determine a quantum mechanical potential completely. For example, consider the ideal case $n \to \infty$ where the number of data goes to infinity for a discrete x–variable (so derivatives with respect to x have to be understood as differences) at zero temperature ($\beta \to \infty$). This, at least, would allow to obtain $p(x|\hat{x}, v) = |\phi_0(x)|^2$ to any desired precision. But even when we restrict ourselves to the case of a local potential, we would also need, for example, the ground state energy E_0 and $\phi_0^*(x)\phi_0''(x)$ to determine $v(x)$ from the eigenvalue equation of H

$$v(x) = E_0 + \frac{1}{2m} \frac{\phi_0^*(x)\phi_0''(x)}{|\phi_0(x)|^2}, \tag{7.22}$$

where $\phi_0'' = \partial^2 \phi_0(x)/\partial x^2$ (or a discretized version thereof). For a finite number of data, a nonlocal potential, continuous x, or finite temperature the situation is obviously even worse. In the high temperature limit, for example, $p(x|\hat{x}, v)$ becomes uniform and independent of the potential. Summarizing, even in the ideal case where the complete true likelihood $p(x|\hat{x}, v)$ is assumed to be known, the problem of reconstructing the potential can still be ill–posed. (The corresponding time–dependent problem, i.e., the reconstruction of a potential v given the complete time–dependent likelihood, is treated in [540]. A Bayesian approach for time–dependent systems, based on a finite number of data, is discussed in Section 7.2.) Hence, while *a priori* information is crucial for every learning problem [291,294,295], the reconstruction of a quantum mechanical potential is particularly sensitive to the implemented *a priori* information.

As next step we include, in addition to the likelihood terms, *a priori* information in form of a prior $p(v|D_0)$ (For the sake of simplicity we will in the following not denote D_0 explicitly and write $p(v)$ for the prior.) Having specified $p(v)$, a Bayesian approach aims at calculating the predictive density

defined in Eq. (2.64), $p(x|O,D,D_0) = \int dv\, p(x|O,v)p(v|D,D_0)$. Within a maximum posterior approximation the functional integral over potentials v in the predictive density (2.64) can be calculated by Monte Carlo methods or, as done in the following, in saddle point approximation by selecting the potential with maximal posterior. The posterior density of v is according to Bayes rule (2.58) proportional to the product of training likelihood and prior

$$p(v|D) \propto p(v) \prod_i < |\phi(x_i)|^2 > . \tag{7.23}$$

Hence the Maximum Likelihood Approximation we have discussed in (the last) Section 7.1.1 is equivalent to a Maximum A Posteriori Approximation under the assumption of a uniform prior.

A Gaussian process prior for v reads, (writing the factor λ separately)

$$p(v) = \left(\det \frac{\lambda}{2\pi}\mathbf{K}_0\right)^{\frac{1}{2}} e^{-\frac{\lambda}{2}\langle v-v_0|\mathbf{K}_0|v-v_0\rangle}, \tag{7.24}$$

with mean v_0, representing a reference potential or template for v, and real–symmetric, positive (semi–)definite covariance operator $(1/\lambda)\mathbf{K}_0^{-1}$, acting on potentials v rather than on wave functions ϕ_α. The operator $\mathbf{K}_0$ defines a scalar product and thus a distance measuring the deviation of v from v_0. The most common priors are *smoothness priors* where $\mathbf{K}_0$ is taken as differential operator (defining a Sobolev distance). Examples of smoothness related inverse prior covariances are the negative Laplacian $\mathbf{K}_0 = -\Delta$ which we have already met in Eq. (2.108), or operators with higher derivatives like a Radial Basis Function prior (2.109) with pseudo-differential operator $\mathbf{K}_0 = \exp(-\sigma_0^2\Delta/2)$ [8,162,183,263,453,507,519].

Like for classical problems the prior density can be parameterized by hyperparameters to become more flexible. As discussed in Chapter 5, hyperparameters are in a Bayesian calculation included as integration variables in the predictive density (2.64) or, in MAP, in the maximization of the posterior [33,52,291]. Hyperparameters allow to generalize the (zero–dimensional) point–maxima of Gaussian priors (with invertible covariance) to a higher dimensional space of optimal solutions. We have seen that for a Gaussian process prior the mean or reference potential v_0 and the covariance $\mathbf{K}_0^{-1}/\lambda$ can in this way be adapted to the data.

To be more general let us consider *a priori* information formulated in terms of some *approximate invariance* or *approximate symmetry* [295]. In contrast to an exact invariance where it is sufficient to restrict v to be invariant under some transformation, it has been shown in Section 3.6.1 that implementing approximate invariance requires a distance measure which quantifies the deviation from exact invariance. To this end, consider a unitary symmetry operation S with $S^\dagger = S^{-1}$, $S^\dagger$ denoting the hermitian conjugate of S. Further, define an operator $\mathbf{S}$ acting on (local or nonlocal) potentials V, by $\mathbf{S}V$

$= S^\dagger V S$. In case of an exact symmetry, V commutes with S, $[V, S] = 0$, and thus $\mathbf{S}V = S^\dagger V S = V$. In case of an approximate symmetry we may choose a prior

$$p(V) \propto e^{-E_S} \tag{7.25}$$

with 'symmetry energy' or 'symmetry error'

$$E_S = \frac{1}{2} \langle V - \mathbf{S}V \,|\mathbf{K}_S|\, V - \mathbf{S}V \rangle = \frac{1}{2} \langle V \,|\mathbf{K}_0|\, V \rangle , \tag{7.26}$$

some positive (semi–)definite $\mathbf{K}_S$ and hence positive semi–definite $\mathbf{K}_0 = (\mathbf{I} - \mathbf{S})^\dagger \mathbf{K}_S (\mathbf{I} - \mathbf{S})$, $\mathbf{I}$ denoting the identity. Symmetric V are within the null space of $\mathbf{K}_S$. If $\mathbf{S}$ belongs to a Lie group it can be expressed by a Lie group parameter θ and generator $\mathbf{s}$ of the corresponding infinitesimal symmetry operation as $\mathbf{S}(\theta) = \exp(\theta \mathbf{s})$. Hence we can define an error E_s with respect to the infinitesimal operation $\mathbf{s}$ with $\mathbf{K}_S = \mathbf{I}$,

$$E_s = \lim_{\theta \to 0} \frac{1}{2} \left\langle \frac{V - \mathbf{S}(\theta)V}{\theta} \middle| \frac{V - \mathbf{S}(\theta)V}{\theta} \right\rangle = \frac{1}{2} \langle V \,|\mathbf{s}^\dagger \mathbf{s}|\, V \rangle . \tag{7.27}$$

Choosing, for instance, $\mathbf{s}$ as the derivative operator (for vanishing or periodic boundary terms) results in the typical Laplacian smoothness prior which measures the degree of symmetry of v under infinitesimal translations.

Another possibility to implement approximate symmetries is given by a prior with symmetric reference potential $V_S = \mathbf{S}V_S$

$$E_{V_S} = \frac{1}{2} \langle V - V_S | V - V_S \rangle . \tag{7.28}$$

In contrast to Eq. (7.26) which is minimized by any symmetric V, this term is minimized only by $V = V_S$. Note that in Eq. (7.26) an explicit non–zero reference potential V_0 can also be included, meaning that not V but the difference $V - V_0$ is expected to be approximately symmetric.

Finally, a certain deviation κ from exact symmetry might be expected. This can be implemented by including 'generalized data terms' [291]

$$E_{S,\kappa} = \frac{1}{2} |E_S(V) - \kappa|^2 = \frac{1}{2} |\frac{1}{2} \langle V - \mathbf{S}V | V - \mathbf{S}V \rangle - \kappa|^2 , \tag{7.29}$$

similar to the usual mean squared error terms used in regression.

Stochastic process priors have, compared to priors over parameters ξ, the advantage of implementing *a priori* knowledge explicitly in terms of the function values $v(x)$. Gaussian processes always correspond to simple quadratic error surfaces. While technically very convenient, they imply, on the other hand, a strong restriction. Arbitrary prior processes, however, can easily be built by using mixtures of Gaussian processes without loosing the advantage of an explicit prior implementation [291, 294, 295, 297]. (We want to point

out that using a mixture of Gaussian process priors does not restrict v to a mixture of Gaussians.)

A mixture of M Gaussian processes with component means v_k and inverse component covariances $\lambda \mathbf{K}_k$ reads

$$p(v) = \sum_k^M p(v,k) = \sum_k^M p(k)\, p(v|k) = \sum_k^M \frac{p(k)}{Z_k} e^{-\frac{\lambda}{2}\langle v-v_k|\mathbf{K}_k|v-v_k\rangle}, \quad (7.30)$$

with $Z_k = \left(\det \frac{\lambda}{2\pi}\mathbf{K}_k\right)^{-\frac{1}{2}}$ and mixture probabilities $p(k)$. The parameter λ plays the role of an inverse mixture temperature. Analogous to annealing techniques changing λ allows to control the 'smoothness' of the mixture [294, 297].

Using a standard Gaussian smoothness prior as in Eq. (7.24) with zero reference potential $v_0 \equiv 0$ (and zero boundary conditions for v) leads to flat potentials for large smoothness factors λ. Especially in such cases it turned out to be useful to include besides smoothness also *a priori* information which determines the depth of the potential. One such possibility is to include information about the *average energy*

$$U = \sum_\alpha p_\alpha E_\alpha = < E > . \quad (7.31)$$

We may remark, that for fixed boundary values of v the value of the average energy can not be controlled by simply adding a constant to the potential. The average energy can, however, be set to a value κ by including a Lagrange multiplier term

$$E_U = \mu(U - \kappa). \quad (7.32)$$

Similarly, and technically sometimes easier, one can include noisy 'energy data' of the form

$$p_U \propto e^{-E_U}, \quad E_U = \frac{\mu}{2}(U-\kappa)^2. \quad (7.33)$$

In the limit $\mu \to \infty$ this means $U \to \kappa$, i.e., the average energy U corresponding to a solution v is set to κ, and both approaches coincide.

Let us consider prior densities being a product of a Gaussian prior p as in Eq. (7.24), or, more generally, a mixture of Gaussian processes as in Eq. (7.30), and a non–Gaussian energy prior p_U of the form of Eq. (7.33). In that case, the stationarity equation we have to solve to maximize the posterior density of Eq. (7.23) reads

$$0 = \delta_{v(x)} \ln p(v) + \delta_{v(x)} \ln p_U(v) + \sum_i \delta_{v(x)} \ln p(x_i|\hat{x}, v). \quad (7.34)$$

While $\delta_{v(x)} p(x_i|\hat{x}, v)$ has already been calculated in Section 7.1.1, we now need also $\delta_{v(x)} p(v)$ and $\delta_{v(x)} p_U(v)$. For a Gaussian $p(v)$ the functional derivative is

easily found to be

$$\delta_v \ln p(v) = \frac{\delta_v p(v)}{p(v)} = -\lambda \mathbf{K}_0 (v - v_0). \tag{7.35}$$

Similarly, for a mixture of Gaussian processes it is not difficult to show that

$$\delta_v \ln p(v) = -\lambda \sum_k^M p(k|v) \mathbf{K}_k (v - v_k), \tag{7.36}$$

where $p(k|v) = p(v,k)/p(v)$.

To get the functional derivative of the non–Gaussian p_U in Eq. (7.33) we calculate first

$$\delta_{v(x)} U = < \delta_{v(x)} E > -\beta < E \delta_{v(x)} E > + \beta < E >< \delta_{v(x)} E > . \tag{7.37}$$

With $\delta_{v(x)} E_\alpha$ from Eq. (7.20), this yields

$$\delta_{v(x)} E_U = \mu (U - \kappa) \left(< |\phi(x)|^2 > -\beta \left(< E\, |\phi(x)|^2 > - U < |\phi(x)|^2 > \right) \right). \tag{7.38}$$

Collecting all terms, we can now solve the stationarity equation (7.34) by iteration

$$v^{(r+1)} = v^{(r)} + \eta \mathbf{A}^{-1} \left(\lambda \mathbf{K}_0 (v_0 - v^{(r)}) + \sum_i \delta_v \ln p(x_i|\hat{x}, v^{(r)}) - \delta_v E_U(v^{(r)}) \right), \tag{7.39}$$

where we introduced $\mathbf{K}_0 = \sum_k p(k|v)\, \mathbf{K}_k$ and $v_0 = \mathbf{K}_0^{-1} \sum_k p(k|v)\, \mathbf{K}_k v_k$, a step width η, and positive definite iteration matrix $\mathbf{A}$ has to be selected (see Appendix C). Choosing $\mathbf{A}$ as the identity matrix means moving in the direction of the gradient of the posterior. Taking for $\mathbf{A}$ the Hessian, one obtains the Newton method. Quasi–Newton methods, like the DFP (Davidon–Fletcher–Powell) or BFGS (Broyden–Fletcher–Goldfarb–Shanno) variable metric methods, approximate the Hessian iteratively [28,40,134,398,410]. For the case of solving for continuous functions see [7].

A simple and useful choice in our case is $\mathbf{A} = \lambda \mathbf{K}_0$ which approximates the Hessian. For a single Gaussian prior this choice does not depend on v and has thus not to be recalculated during iteration. Eq. (7.39) then becomes

$$v^{\text{new}} = (1 - \eta) v^{\text{old}} + \eta \left(v_0 + (\lambda \mathbf{K}_0)^{-1} \Big(\sum_i \delta_v \ln p(x_i|\hat{x}, v^{\text{old}}) - \delta_v E_U^{\text{old}} \Big) \right). \tag{7.40}$$

Due to the nonparametric approach for the potential combined with *a priori* information implemented as stochastic process, the Bayesian approach formulated in the previous sections is clearly computationally demanding.

The situation for inverse quantum theory is worse than, e.g., for Gaussian process priors in regression problems (i.e., for a Gaussian likelihood, local in the regression function) where it is sufficient to work with matrices having a dimension equal to the number of training data [507, 522]. In our case, where the likelihood is nonlocal in the potential and also non–Gaussian prior terms may occur, the stationarity equation has to be solved by discretizing the problem.

The following section demonstrates that at least one–dimensional problems can be solved numerically without further approximation. Higher dimensional problems, however, like many–body systems, require additional approximations. In such higher dimensional situations, the potential may be parameterized (without skipping necessarily the prior terms) or the problem has to be divided into lower dimensional subproblems, for example by restricting v to certain (typically, additive or multiplicative) combinations of lower dimensional functions. (Like it is done in additive models [203], projection pursuit [227], or neural network like [52] approaches.)

7.1.3 Numerical case study: Inverse quantum statistics

As numerical example we study the reconstruction of a two–body potential by measuring inter–particle distances x_r. Consider the two–body problem

$$\frac{P_1^2}{2m_1} + \frac{P_2^2}{2m_2} + v(x_1 - x_2)\delta(x_1 - x_2 - x_1' + x_2')\delta(x_1 + x_2 - x_1' - x_2') \quad (7.41)$$

with single particle momenta $P_i = -i\partial/\partial x_i$. The problem is transformed to a one–body problem in the relative coordinates in the usual way by introducing $x_r = x_1 - x_2$, $P_r = (m_1 P_1 - m_2 P_2)/(m_1 + m_2)$, $x_c = (m_1 x_1 + m_2 x_2)/(m_1 + m_2)$, $P_c = P_1 + P_2$, $m = (m_1 m_2)/(m_1 + m_2)$, and $M = m_1 + m_2$, resulting in an eigenvalue equation for wave functions ψ_α in relative coordinates x_r

$$\left(\frac{P_r^2}{2m} + v(x_r)\right)\psi_\alpha(x_r) = E_\alpha \psi_\alpha(x_r). \quad (7.42)$$

The total energy is additive $E_\alpha^{\text{total}}(P_c) = E_\alpha + P_c^2/(2M)$ and the thermal probabilities p^{total} factorize. Integrating out the center of mass motion leaves $p_\alpha = e^{-\beta E_\alpha}/Z$, with E_α being the eigenvalues of Eq. (7.42).

Figs. 7.1 – 7.3 show typical results for the numerical reconstruction of a one–dimensional, strictly symmetric potential, fulfilling $v(x) = v(-x)$ and set to zero at the boundaries. Training data have been sampled from a 'true' likelihood function (thin lines), resulting in an empirical density $P_{\text{emp}}(x) = n(x)/n$ (shown by bars), where $n(x)$ denotes the number of times the value x occurs in the training data. The 'true' likelihood has been constructed from a 'true' potential (thin lines) choosing periodic boundary conditions for the

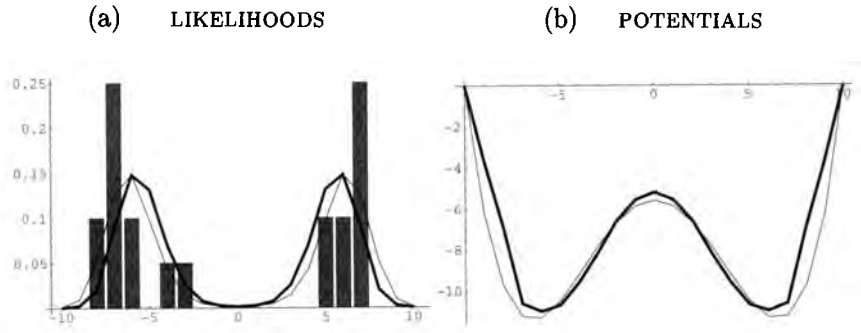


Figure 7.1: BIQS 1: Approximation of symmetric potentials 1. Shown are likelihoods (left) and potentials (right): Original likelihood and potential (thin lines), approximate likelihood and potential (thick lines), empirical density (bars). The parameters used are: 20 data points for a particle with $m = 0.1$, truncated RBF covariances as in Eq. (7.43) with $\sigma_0 = 7$, $\lambda = 0.001$, energy penalty term E_U with $\mu = 20$ and reference value $\kappa = -9.66 = U(v_{\text{true}})$, average energy $U(v) = -9.33$ for the approximate v, ground state energy $E_0(v) = -9.52$, inverse physical temperature $\beta = 1$, and a potential fulfilling $v(x) = v(-x)$ and $v = 0$ at the boundaries.

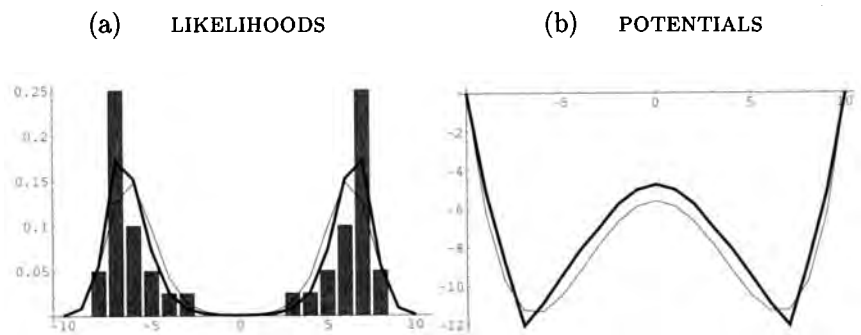


Figure 7.2: BIQS 2: Approximation of symmetric potentials 2. Same data and parameters as for Fig. 7.1 with the exception of $\sigma_0 = 4$, which means a weaker smoothness constraint, and $\mu = 5$. ($U(v) = -9.46$.) To allow an easier comparison with the reconstructed likelihood the figure shows the symmetrized empirical density $P_{\text{sym}} = (P_{\text{emp}}(x) + P_{\text{emp}}(-x))/2$.

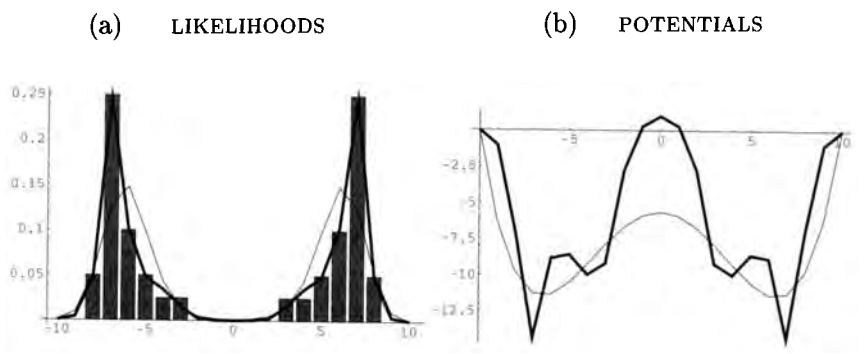


Figure 7.3: BIQS 3: Approximation of symmetric potentials 3 (overfitting). Same data and parameters as for Fig. 7.2 but with even weaker smoothness constraint $\sigma_0 = 1$. ($\mu = 5$, empirical density symmetrized, $U(v) = -9.59$.) Compared with Figs. 7.1 and 7.2, the empirical density (bars on the left) is better approximated, but not the original likelihood function (thin line on the left) and the potential (thin line on the right).

wavefunctions. A zero reference potential $v_0 \equiv 0$ and a truncated Radial Basis Function (RBF) prior as in (3.285) has been used

$$\mathbf{K}_0 = \sum_{k=0}^{3} \frac{\sigma_0^{2k}}{k! 2^k} (-1)^k \Delta^k, \tag{7.43}$$

Δ^k denoting the kth iterated Laplacian, which includes higher derivatives, producing a smoother reconstructed potential (cf. Section 7.1.2). The approximate potentials (thick lines) have been obtained by iterating Eq. (7.39), including a term E_U adapting the thermal energy average to that of the original potential. As iteration matrix we used $\mathbf{A} = \lambda \mathbf{K}_0$ together with an adaptive step size η. An initial guess for the potential has been obtained by adding negative δ–peaks on the data points (except for data on the boundary), i.e., $v^{(0)} = -\sum_i \delta_{x,x_i}$. The number of iterations necessary to obtain convergence has been typically between 50 and 100.

Comparing Figs. 7.1 – 7.3 one sees that less smoothness leads to a better fit of the empirical density. Larger smoothness, on the other hand, leads to a better fit in regions where smoothness is an adequate prior. Near the boundaries, however, where the original potential is relatively steep, a higher smoothness leads to a poorer approximation. A remedy would be, for example, an adapted reference potential v_0.

Fig. 7.4 presents an application of a mixture of Gaussian process priors as given in Eq. (7.30). Such mixture priors can in principle be used to construct an arbitrary prior density, adapted to the situation under study. For the

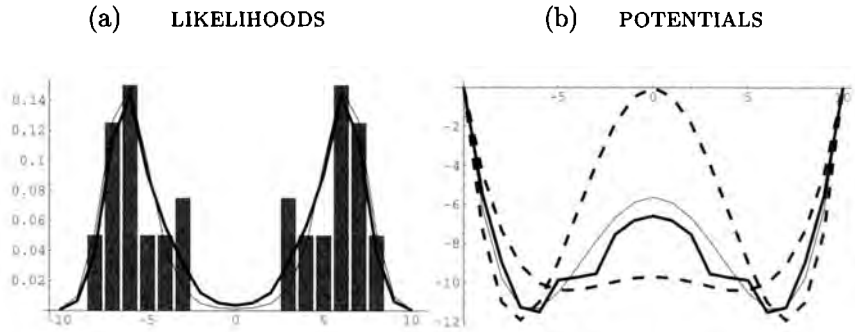


Figure 7.4: BIQS 4: Approximation of symmetric potentials 4 (Gaussian mixture prior). Shown is the approximation of a symmetric potential using a mixture of Gaussian process priors. The left hand side shows likelihoods and the right hand side potentials: Original likelihood and potential (thin lines), approximate likelihood and potential (thick lines), symmetrized empirical density (bars), and the two reference potentials v_1,v_2 (dashed, v_2 deeper in the middle). The parameters used are: 20 data points for a particle with $m = 0.1$, inverse physical temperature $\beta = 1$, $\mathbf{K}_0 = -\Delta$, inverse mixture temperature $\lambda = 0.1$, energy penalty factor $\mu = 10$ for average energy $\kappa = -9.66 = U(v_{\text{true}})$ (and $U(v) = -9.55$, $E_0(v) = -9.82$), $v(x) = v(-x)$ symmetric, and $v = 0$ at the boundaries. Because the data support both reference potentials v_1 and v_2, the approximate v is in regions with no data essentially a smoothed mixture between v_1 and v_2 with mixture coefficients for prior components $p(1|v) = 0.3$, $p(2|v) = 0.7$.

numerical example a two component mixture has been chosen with equal inverse component covariances $\mathbf{K}_k = \mathbf{K}_0$ of the form of Eq. (7.43) and two reference potentials v_i (shown as dashed lines) with the same average energy U. In the special situation shown in the figure both reference functions v_i fit similarly well to the empirical data. (The final mixture coefficients for v_1 and v_2 are $p(1|v) = 0.3$ and $p(2|v) = 0.7$.) Hence in the no–data region the approximate potential v becomes a smoothed, weighted average of v_1 and v_2. Because both reference potentials also coincide relatively well with the original v near the boundaries, the approximation in Fig. 7.4 is better than those in Figs. 7.1 - 7.3.

In conclusion, the two one–dimensional examples show that a direct numerical solution of the presented Bayesian approach to inverse quantum theory is feasible. The Bayesian approach to inverse quantum statistics is summarized in Fig. 7.5.

7.1.4 Classical approximation

Before discussing a possible approximation to many–body systems we will first study the classical limit of inverse quantum statistics. The classical limit is much easier to solve than the full quantum mechanical problem and may for large masses, for example, already give a useful approximation.

The phase space density of a classical canonical ensemble is given by

$$p(x, p_{\mathrm{cl}}|v) = Z^{-1} e^{-\beta\left(\frac{p_{\mathrm{cl}}^2}{2m} + v(x)\right)}, \tag{7.44}$$

with

$$Z = \int dp_{\mathrm{cl}}\, dx\, e^{-\beta\left(\frac{p_{\mathrm{cl}}^2}{2m} + v(x)\right)}. \tag{7.45}$$

Here we used p_{cl} to denote the classical momentum to distinguish it from a density p. The density $p(x|v)$ for measuring x [to simplify the notation we abstain in this context from denoting the observable O explicitly] is then obtained by integrating over p_{cl},

$$p(x|v) = \int dp_{\mathrm{cl}}\, p(x, p_{\mathrm{cl}}|v) = Z_x^{-1} e^{-\beta v(x)}, \tag{7.46}$$

where

$$Z_x = \int dx\, e^{-\beta v(x)}, \tag{7.47}$$

is the norming constant for a single measurement. Notice that the classical $p(x|v)$ is mass independent and, most important, it can be obtained directly from $v(x)$ without having to solve an eigenvalue problem like in the quantum case.

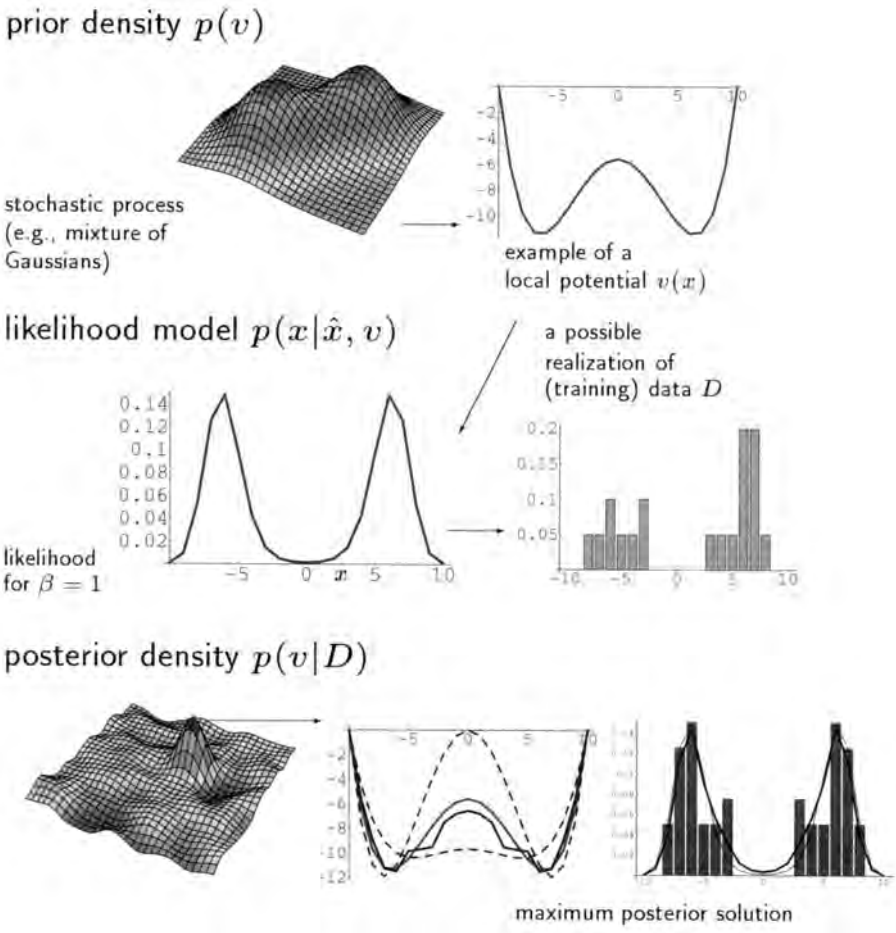


Figure 7.5: BIQS 5: Scheme of Bayesian inverse quantum statistics. The 2–dimensional figures of the prior and posterior densities are symbolic representations of stochastic processes over potentials.

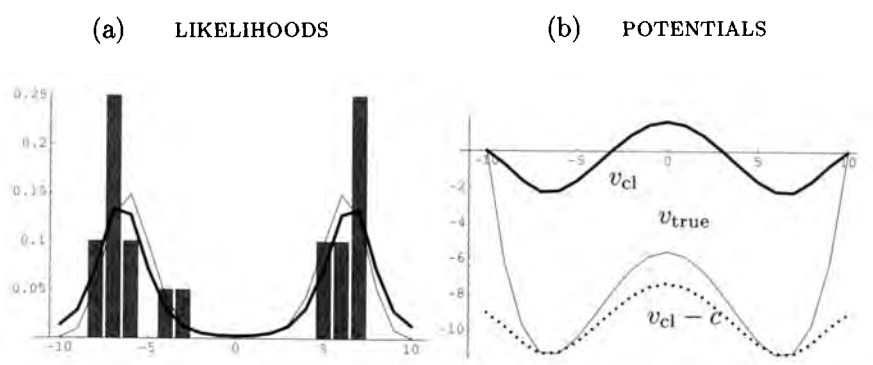


Figure 7.6: BIQS 6: Classical approximation of a symmetric potential. Shown are likelihoods (left) and potentials (right): Original likelihood and potential (thin lines), approximate likelihood and potential $v_{\rm cl}$ (thick lines), empirical density (bars). The dotted line shows $v_{\rm cl} - c$ with constant $c = \min[v_{\rm cl}] - \min[v_{\rm true}]$. Except for the fact that no energy penalty term has been used for this classical calculation, the parameters and data are the same as in Fig. 7.1. (Twenty data points, sampled from the true quantum mechanical likelihood, truncated RBF covariances (7.43) with $\sigma_0 = 7$, $\lambda = 0.001$, inverse physical temperature $\beta = 1$, $v(x) = v(-x)$ and $v = 0$ at the boundaries.)

Analogous to the quantum mechanical approach the classical likelihood model (7.46) for position measurements can now be combined with a prior model for potentials v, leading to a posterior density $p(v|D)$. Adding a Gaussian process prior, the log–posterior becomes

$$\ln p(v|D) = -\beta \sum_{i=1}^{n} v(x_i) - \frac{\lambda}{2} \langle v - v_0 \,|\mathbf{K}_0|\, v - v_0 \rangle - n \ln Z_x - \det \left(\frac{\lambda}{2\pi} \mathbf{K}_0 \right)^{\frac{1}{2}} . \tag{7.48}$$

Again we may refer to a maximum posterior approximation and consider the potential which maximizes the posterior as the solution of our reconstruction problem. The corresponding stationarity equation is found by setting the functional derivative of the log–posterior with respect to $v(x)$ to zero,

$$0 = \delta_v \ln p(v|D) = -\beta N - \lambda \mathbf{K}_0 (v - v_0) + n\beta p(x|v). \tag{7.49}$$

Choosing an initial guess $v^{(0)}$, Eq. (7.49) can be solved by straightforward iteration. The results of a classical calculation (with parameters and data as in Fig. 7.1, but without energy penalty term) are shown in Fig. 7.6.

7.1.5 Numerical case study: Reconstruction of perturbed periodic potentials using hyperfields and auxiliary fields

To demonstrate the implementation of adapted, situation-specific *a priori* information using the methods developed in Sections 3 and 5, this section discusses the inverse problem of reconstructing a perturbed periodic potential of a quantum statistical system [306]. Starting from a single Gaussian prior process, prior implementations of increasing complexity and flexibility will be considered.

It has been shown in Sections 5.3 and 5.4 that already Gaussian prior factors

$$p(v) = \left(\det \frac{1}{2\pi}\mathbf{K}_0\right)^{\frac{1}{2}} e^{-\frac{1}{2}\langle v-v_0|\mathbf{K}_0|v-v_0\rangle}, \tag{7.50}$$

(like Eq. (7.24) or (3.156) with the regularization factor λ absorbed in $\mathbf{K}_0$), i.e., quadratic prior energies

$$E(v) = \frac{1}{2}\langle v - v_0 |\mathbf{K}_0| v - v_0\rangle, \tag{7.51}$$

can implement very specific *a priori* information by adapting covariance and mean function. For example, approximate periodicity can be implemented by choosing a 'periodic' covariance or by choosing a periodic reference potential v_0 as mean function of the Gaussian process.

More flexibility is added by including a global real (mixing) or binary (switching) hyperparameter θ as in Eq. (5.116),

$$E(v|\theta) = \frac{1-\theta}{2}\langle v - v_1 |\mathbf{K}_1| v - v_1\rangle + \frac{\theta}{2}\langle v - v_2 |\mathbf{K}_2| v - v_2\rangle. \tag{7.52}$$

This allows, for example, to switch (or mix) a smoothness prior with an approximate periodicity prior if one is uncertain whether to expect a smooth potential or a potential of approximately periodic structure. Summing over the possible values of an integer hyperparameter θ results in mixture models treated in Section 6.

Many more degrees of freedom enter the formalism if the switching or mixing process is allowed to be local. This is made possible by introducing local hyperfields $\theta(x)$, like in the binary case as in Eq. (5.128)

$$E(v|\theta) = \frac{1}{2}\int dx \left([1-\theta(x)]|\omega_1(x)|^2 + \theta(x)|\omega_2(x)|^2\right) + \ln Z_{\mathcal{V}}(\theta), \tag{7.53}$$

with filtered difference $\omega_i(x) = \int dx'\, \mathbf{W}_i(x,x')[v(x') - v_i(x')]$ as defined in Eq. (5.122), or, similar to Eq. (5.142)

$$E(v|\theta) = \frac{1}{2}\langle v - \tilde{v}_0(\theta)| v - \tilde{v}_0(\theta)\rangle + \frac{1}{2}\langle v |\mathbf{K}_2| v\rangle, \tag{7.54}$$

Gaussian prior	Eq.
$E(v) = \frac{1}{2}\langle v - v_0 \vert \mathbf{K}_0 \vert v - v_0\rangle$	(7.51)
with hyperparameter θ	
$E(v\vert\theta) = \frac{1-\theta}{2}\langle v - v_1 \vert \mathbf{K}_1 \vert v - v_1\rangle$ $+\frac{\theta}{2}\langle v - v_2 \vert \mathbf{K}_2 \vert v - v_2\rangle$	(7.52)
with local hyperfield $\theta(x)$	
$E(v\vert\theta) = \frac{1}{2}\int dx \left([1-\theta(x)]\vert\omega_1(x)\vert^2 + \theta(x)\vert\omega_2(x)\vert^2\right) + \ln Z_{\mathcal{V}}(\theta)$	(7.53)
$E(v\vert\theta) = \frac{1}{2}\langle v - \tilde{v}_0(\theta) \vert v - \tilde{v}_0(\theta)\rangle + \frac{1}{2}\langle v \vert \mathbf{K}_2 \vert v\rangle$	(7.54)
Non–Gaussian prior with auxiliary field $B(x;v)$	
$E(v) = \frac{1}{2}\int dx \left([1-B(x)]\vert\omega_1(x)\vert^2 + B(x)\vert\omega_2(x)\vert^2\right)$	(7.55)

Table 7.2: Summary of basic prior energy variants.

where the combination of two prior energy terms corresponds to one v–dependent prior energy term with reference potential $v_0 = (\mathbf{I} + \mathbf{K}_2)^{-1}\tilde{v}_0$ and inverse covariance $\mathbf{K}_0 = \mathbf{I} + \mathbf{K}_2$. Such priors allow one, for example, to switch off a periodicity prior locally where (according to the training data) a periodic potential is expected to be perturbed.

Alternatively to introducing hyperfields as new degrees of freedom the local switching or mixing process can be governed by auxiliary fields $B(x;v)$ expressed in terms of the potential v. The resulting priors are, as in Eq. (5.155), explicitly non–Gaussian in v, e.g., with binary B,

$$E(v) = \frac{1}{2}\int dx \left([1-B(x)]\vert\omega_1(x)\vert^2 + B(x)\vert\omega_2(x)\vert^2\right). \tag{7.55}$$

Table 7.2 summarizes the prior energy variants discussed in this section.

Auxiliary fields defined in terms of the potential v contribute to the functional derivative with respect to v. As an example consider a prior where two filtered differences are mixed by an auxiliary field $B(x)$ and an additional prior factor $p(B)$ is included, for example, to prevent fast oscillations of $B(x)$. With $B(x) = \sigma(u(x) - \vartheta)$, threshold ϑ, sigmoidal function $\sigma(\cdot)$ as in Eq. (5.130), and $u(x) = \vert\omega_1(x)\vert^2 - \vert\omega_2(x)\vert^2$ this gives

$$p(v) \propto e^{-\frac{1}{2}\int dx \left\vert[1-B(x)]\omega_1(x)+B(x)\omega_2(x)\right\vert^2 - E_B}. \tag{7.56}$$

Analogous to Eq. (5.154), the term

$$E_B = \int dx\, E_B(x) \tag{7.57}$$

represents an auxiliary prior energy formulated in terms of the mixing function $B(x)$. Like $\omega(x)$ the function value $E_B(x)$ may depend on the whole function B rather than only on the function value $B(x)$. Using $\omega_i(x) = \langle x|\, \mathbf{W}_i(v - v_i)\rangle$ we find

$$\delta_{v(x)}\omega_i(x') = \mathbf{W}_i(x', x), \tag{7.58}$$

and thus

$$\delta_{v(x)}u(x') = 2\left(\mathbf{W}_1^T(x, x')\,\omega_1(x') - \mathbf{W}_2^T(x, x')\,\omega_2(x')\right). \tag{7.59}$$

Furthermore, we obtain for the functional derivative of E_B

$$\delta_{v(x)}E_B(x') = \int dx''\ \left[\delta_{v(x)}B(x'')\right]\ \left[\delta_{B(x'')}E_B(x')\right], \tag{7.60}$$

where with Eq. (5.150)

$$\delta_{v(x)}B(x'') = \sigma'(u(x'') - \vartheta)\delta_{v(x)}u(x''), \tag{7.61}$$

and $\sigma'(u) = d\sigma(u)/du$. For a prior energy as in (5.157) with local components quadratic in $B(x)$,

$$E_B(x) = |\omega_B(x)|^2, \tag{7.62}$$

$\omega_B(x)$ defined in Eq. (5.160), the functional derivative with respect to $B(x)$ becomes

$$\delta_{B(x)}E_B(x') = 2\mathbf{W}_B^T(x, x')\,\omega_B(x'). \tag{7.63}$$

For a non–Gaussian prior with energy (5.158) an additional derivative of the sigmoid appears. Now all terms can be collected and inserted into the functional derivative of the prior (7.56)

$$\begin{aligned}\delta_v \ln p(v) \;=\; -\int dx \Big(&[[1 - B(x)]\omega_1(x) + B(x)\omega_2(x)] \\ &\times \big([1 - B(x)]\delta_v\omega_1(x) + B(x)\delta_v\omega_2(x) \\ &\quad + \delta_v B(x)[\omega_2(x) - \omega_1(x)]\big) + \delta_v E_B(x)\Big). \end{aligned} \tag{7.64}$$

As numerical application of Bayesian Inverse Quantum Theory (BIQT) and to test several variants of implementing *a priori* information we will study the reconstruction of an approximately periodic, one–dimensional potential. Such a potential may represent a one–dimensional surface where the periodic structure of a regular crystal is distorted by impurities, located at unknown positions and of unknown form.

To test the quality of reconstruction algorithms, artificial data will be sampled from a model with known 'true' potential v_{true}. Selecting a specific prior model and applying the corresponding Bayesian reconstruction algorithm to the sampled data, we will be able to compare the reconstructed potential with the original one. In particular, we will take as true potential the following perturbed periodic potential

$$v_{\text{true}}(x) = \begin{cases} \sin\left(\frac{2\pi}{6}x\right); & 1 \leq x \leq 12,\ 25 \leq x \leq 36, \\ \sin\left(\frac{2\pi}{12}x\right); & 13 \leq x \leq 24, \end{cases} \tag{7.65}$$

using for the numerical calculations a mesh of size 36. Considering a system prepared as a canonical ensemble, the potential v_{true} defines a corresponding canonical density operator ρ as given in Eq. (7.3). Artificial data D can then be sampled according to the likelihood model of quantum mechanics (7.1). For the following examples, $n = 200$ data points representing position measurements have been sampled using the transformation method [410]. In all calculations we used periodic boundary conditions for quantum mechanical wave functions while the potential v has been set to zero at the boundaries.

We will now discuss the results of a Bayesian reconstruction under varying prior models. As the first example, consider a simple Gaussian prior (7.50) with negative Laplacian inverse covariance $\mathbf{K}_0 = -\lambda\Delta$, zero reference potential $v_0 \equiv 0$, and an additional prior factor (7.33) representing a noisy measurement of the average energy. The reconstruction results are shown in Fig. 7.7. In particular, the figure on top compares the reconstructed likelihood $p_{\text{BIQT}}(x|\hat{x}, v_{\text{BIQT}})$ with the true likelihood $p_{\text{true}}(x|\hat{x}, v_{\text{true}})$ and with the empirical density, i.e., the relative frequencies of the sampled data $p_{\text{emp}}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)$. Similarly, the lower figure compares the reconstructed potential v_{BIQT} with the true potential v_{true}. Since information on the average energy was available the depth of the potential is well approximated at least at one of its minima. This is sufficient to fulfill the noisy average energy condition. However, because only smoothness and no periodicity information is implemented by the prior the reconstructed potential is too flat. The effect is stronger near the maxima than near the minima of the potential, because near the maxima only few data points are available and hence the reconstructed potential is there dominated by the zero reference potential in the smoothness prior.

To include information on approximate periodicity in the next example the zero reference potential $v_0 \equiv 0$ has been replaced by the strictly periodic reference potential

$$v_0(x) = \sin\left(\frac{2\pi}{6}x\right), \tag{7.66}$$

shown as dashed line in the following figures of potentials. A reconstruction with the periodic reference potential (7.66) but without average energy information, starting the iteration with the reference potential as initial guess $v^{(0)}$

$= v_0$, is shown in Fig. 7.8. Due to missing average energy information the depth of the potential is not well approximated. It is also clearly visible in Fig. 7.8 that the smoothness prior does not favor solutions which are similar to the reference v_0 itself but solutions which have a derivative similar to that of v_0. Fig. 7.8 also displays that the reconstruction of the potential does clearly identify the impurity. As the reference potential is not adapted to the impurity region the reconstruction is there poorer than in the regular region.

Furthermore, it is worth emphasizing that the reconstructed likelihood fits the empirical density well, even slightly better than the true likelihood does. This is due to the flexibility of a nonparametric approach which allows to fit the fluctuations of the empirical density caused by the finite sample size. The effect is well known in empirical learning and leads to overfitting if the influence of the prior becomes too small. Since observational data influence the reconstruction only through the likelihood, the reconstruction of potentials is in general a more difficult task than the reconstruction of likelihoods. This indicates the special importance of *a priori* information when reconstructing potentials. Indeed, even if the complete likelihood is given, the problem of determining the potential can still be ill–defined in regions where the likelihood is small [540].

A prior model with periodic reference potential can be made more flexible by adapting amplitude, frequency, and phase of the reference potential (7.66). For this purpose one can introduce a hyperparameter vector $\theta = (\theta_1, \theta_2, \theta_3)$ parameterizing amplitude, frequency, and phase and take as reference potential

$$v_0(x;\theta) = \theta_1 \sin\left(\frac{2\pi}{\theta_2}x + \theta_3\right). \tag{7.67}$$

The corresponding maximization of the posterior with respect to θ is easy in that case and does not change the results of Fig. 7.8 where the hyperparameters are already optimally adapted.

Including an additional noisy energy measurement as in Eq. (7.33), Fig. 7.9 shows that the depth of the potential is indeed better approximated than in Fig. 7.8. To avoid local maxima of the posterior, the solution of Fig. 7.8 has been used as the initial guess and the factor μ multiplying the average energy term has been slowly increased to its final value. Fig. 7.9 still represents only a local and no global maximum of the posterior, as can be seen by starting with a different initial guess $v^{(0)}$. In Fig. 7.10 a better solution for the same parameters is presented where the initial guess has been selected using *a priori* information about the location of the impurity region.

An alternative to a Gaussian prior with periodic reference, approximate periodicity can be enforced by the inverse covariance of a Gaussian prior. In this case the prior favors periodicity but no special form of the potential. The prior is thus less specific than a prior with explicit periodic reference function.

Corresponding BIQT results for the inverse covariance

$$\mathbf{K}_0 = -\lambda(\Delta + \gamma\Delta_\theta), \tag{7.68}$$

with weighting factors λ and γ [cf. Eq. (3.145)] are shown in Fig. 7.11. Indeed while the potential is well approximated in regions where many observations have been collected, it is not as well approximated in regions where no or only few data are available. These are the regions where the prior dominates the observational data. In particular, in the case presented in Fig. 7.11, the zero reference function $v_0 \equiv 0$ of an additional Laplacian smoothness prior implements a tendency to flat potentials.

If impurities are expected, a prior with one fixed periodic reference potential for the whole region is no adequate choice. Near impurities one would like to switch off the standard periodic reference potential which in these regions will be misleading. Because it is usually not known in advance where a given reference should be used and where not, those regions must be identified during learning. As a first example we study a prior energy being a local version of Eq. (7.54),

$$E(v) = \frac{\lambda_1}{2}\int dx\,|v(x) - v_0(x)|^2[1 - B(x)] - \frac{\lambda_2}{2}\langle v\,|\Delta|\,v\rangle\,, \tag{7.69}$$

where the first term allows to switch off a given reference locally by means of a binary switching function defined as

$$B(x) = \Theta\left(|v(x) - v_0(x)|^2 - \vartheta\right). \tag{7.70}$$

(An average energy term $E_U = \frac{\mu}{2}(U - \kappa)^2$ could easily be included.) In the prior energy (7.69) the reference v_0 is only used if $|v(x) - v_0(x)|^2$ is smaller than the given threshold ϑ. Starting with a smoothed version of Eq. (7.69) with a real mixing function $B(x) = \sigma\left(|v(x) - v_0(x)|^2 - \vartheta\right)$, the results of Fig. 7.12 have been obtained by changing during iteration $\sigma(x)$ slowly from a sigmoid to a step function. Using a step function for B directly from the beginning leads to nearly indistinguishable results. Compared to Fig. 7.11 the reconstruction in Fig. 7.12 is improved mainly in the unperturbed region where the algorithm can now use the correct reference potential. An additional advantage is that the final auxiliary field $B(x)$ directly shows the identified impurity regions. One sees in Fig. 7.12 that the auxiliary field $B(x)$ is always switched off if the solution $v(x)$ is similar enough to the template $v_0(x)$.

The two v–dependent terms in Eq. (7.69) can be combined [cf. Eqs. (5.142) and (5.143)]. Skipping a term which only depends on v through $B(x)$, one arrives at another prior which also implements local switching. More generally, choosing the prior energy (7.55) for switching between two filtered differences

with two reference potentials v_1 and v_2 leads to

$$E(v) = \frac{\lambda_1}{2}\int dx\,[1-B(x)]|\omega_1(x)|^2 + \frac{\lambda_2}{2}\int dx\,B(x)|\omega_2(x)|^2, \tag{7.71}$$

where the switching is controlled by the binary function

$$B(x) = \Theta\left(|\omega_1(x)|^2 - |\omega_2(x)|^2 - \vartheta\right), \tag{7.72}$$

defined in terms of the filtered differences $\omega_i(x) = (\partial/\partial x)[v(x) - v_i(x)]$. A prior energy (7.71) with two different nonzero reference potentials v_1 and v_2 is obtained, for example, when different nonzero reference potentials are given for the unperturbed and the perturbed region. The number of changes in the switching function $B(x) = \Theta\left(|\omega_1(x)|^2 - |\omega_2(x)|^2 - \vartheta\right)$ can be controlled by adding a prior term $p(B)$ penalizing the number of times the function $B(x)$ changes its value. To avoid local minima for binary $B(x)$, simulated annealing techniques are useful. An initial guess for v, and thus for $B(x)$, has been obtained by writing $v(x) = [1-c(x)]v_1(x) + c(x)v_2(x)$ and optimizing the binary function $c(x)$ by simulated annealing with respect to the likelihood and the additional prior $p(B)$. In particular, starting from $c(x) = 0$, new trial functions have been generated by selecting two points x_1, x_2 randomly and exchanging the function values zero and one in between (see Fig. 7.13). A new trial function has been accepted or rejected using the Metropolis rule $p(\text{accept}) = \min[1, \exp(-\beta_{\text{ann}}\Delta E_{\text{ann}})]$ with ΔE_{ann} denoting the difference in the error between actual function and new trial function. In the present case we have $E_{\text{ann}}(v) = \sum_i E(x_i|\hat{x}, v) + E_B(v)$, where $E(x_i|\hat{x}, v) = -\ln p(x_i|\hat{x}, v)$ and $p(B) \propto \exp(-E_B)$. The annealing temperature $1/\beta_{\text{ann}}$ decreases during optimization.

Fig. 7.14 shows the reconstruction results using the following two reference potentials

$$v_1(x) = \frac{2}{3}\sin\left(\frac{2\pi}{6}x\right), \tag{7.73}$$

$$v_2(x) = \sin^2\left(\frac{2\pi}{6}x\right)\text{sign}\left[\sin\left(\frac{2\pi}{6}x\right)\right]. \tag{7.74}$$

Compared to Fig. 7.12 the reconstruction is improved in the perturbed region, where the algorithm can now rely on a useful reference potential.

Finally, the switching function can be introduced as local hyperfield. As an example for a prior with hyperfield, Fig. 7.15 shows the reconstruction with the prior energy

$$E(v,\theta) = \frac{\lambda_1}{2}\langle v - v_0(\theta)|\, v - v_0(\theta)\rangle - \frac{\lambda_2}{2}\langle v\,|\Delta|\,v\rangle - \ln p(\theta), \tag{7.75}$$

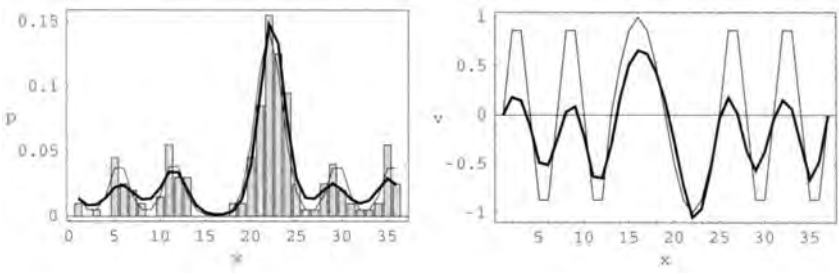


Figure 7.7: BIQS 7: Reconstruction of an approximately periodic potential with smoothness prior. (Gaussian prior with Laplacian inverse covariance, zero reference potential, and additional noisy energy measurement.) Left hand side: Empirical density p_{emp} (bars), true likelihood p_{true} (thin), reconstructed likelihood p_{BIQT} (thick). Right hand side: Reconstructed potential v_{BIQT} (thick) and true potential v_{true} (thin). (With 200 data points, $m = 0.25$ for $\hbar = 1$, $\beta = 4$. Gaussian prior (7.50) with inverse covariance $\mathbf{K}_0 = -\lambda\Delta$, $\lambda = 0.2$, zero reference potential $v_0 \equiv 0$, and an additional energy penalty term of the form (7.33) with $\mu = 1000$ and $\kappa = -0.330$, equal to the true average energy $U(v_{\text{true}})$. The solution has been obtained by iterating according to Eq. (7.39) with $\mathbf{A} = \mathbf{K}_0$, starting with initial guess $v^{(0)} \equiv 0$. The optimal step width η has been determined for each iteration by a line search algorithm.)

where $v_0(x;\theta) = v_1(x)[1 - \theta(x)] + v_2(x)\theta(x)$ with the reference potentials of Eq. (7.73) and Eq. (7.74). A hyperprior $p(\theta)$ has been used penalizing the number of discontinuities of the hyperfield $\theta(x)$, analogous to $p(B)$ for Fig. 7.14. The $E(v|\theta)$ part of the prior energy (7.75) is of the form (7.53) with θ–independent covariances. Hence the θ–independent normalization factor can be skipped. An initial guess for the local hyperfield $\theta(x)$ has been obtained by simulated annealing as described for Fig. 7.14. As in this case optimization is required only with respect to the θ–dependent parts of the posterior, optimizing $\theta(x)$ for given v is faster than optimizing v through $c(x)$ which requires diagonalization of the Hamiltonian H for every new trial function. However, as $\theta(x)$ is independent of v, the hyperfield has to be updated during iteration which has also been done by simulated annealing. As expected, a reconstruction with the non–Gaussian prior corresponding to the prior energy (7.71) is very similar to a reconstruction using hyperfields as in Eq. (7.75).

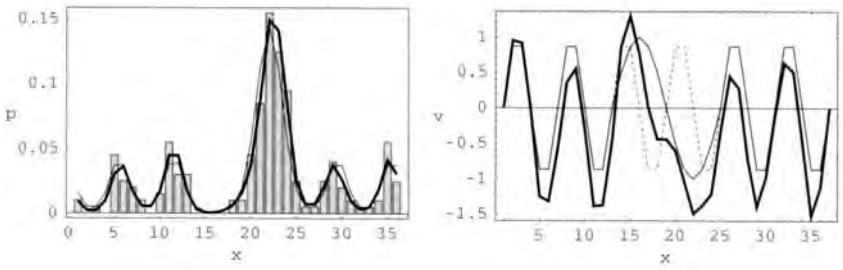


Figure 7.8: BIQS 8: Reconstruction of an approximately periodic potential with periodic reference potential 1. (Gaussian prior with periodic reference potential without noisy energy measurement.) Left hand side: Empirical density p_{emp} (bars), true likelihood p_{true} (thin), reconstructed likelihood p_{BIQT} (thick). Right hand side: Reconstructed potential v_{BIQT} (thick), true potential v_{true} (thin), and reference potential v_0 (dashed) of Eq. (7.66). (Number of data points and parameters m, β, $\mathbf{K}_0$, and λ as for Fig. 7.7 but with $\mu = 0$. The solution has been obtained by iterating according to (7.39) as described for Fig. 7.7 with initial guess $v^{(0)} = v_0$.)

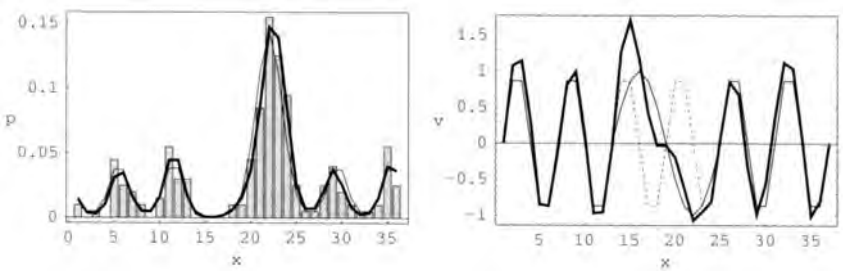


Figure 7.9: BIQS 9: Reconstruction of an approximately periodic potential with periodic reference potential 2. (Gaussian prior with periodic reference potential and additional energy measurement, improving the approximation of the minima.) Reference potential v_0 given in (7.66), energy penalty term as in (7.33) with $\mu = 1000$ and $\kappa = -0.330$. All other parameters as for Fig. 7.8. Iterated with the solution shown in Fig. 7.8 as initial guess $v^{(0)}$.

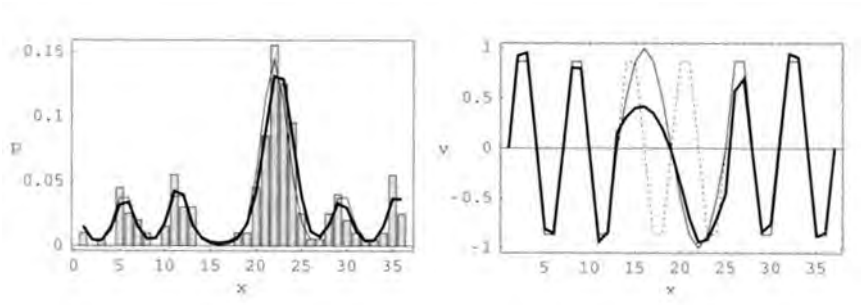


Figure 7.10: BIQS 10: Reconstruction of an approximately periodic potential with periodic reference potential 3. (Gaussian prior with periodic reference potential and additional energy measurement, with initial guess $v^{(0)}$ different from that of Fig. 7.9.) Reference potential v_0 given in (7.66), energy penalty term as in (7.33). All parameters as for Fig. 7.9. Iterated with initial guess $v^{(0)}(x) = v_0(x)$ for $0 < x \leq 12$, $25 \leq x$ and $v^{(0)}(x) = 0$ for $13 \leq x \leq 24$.

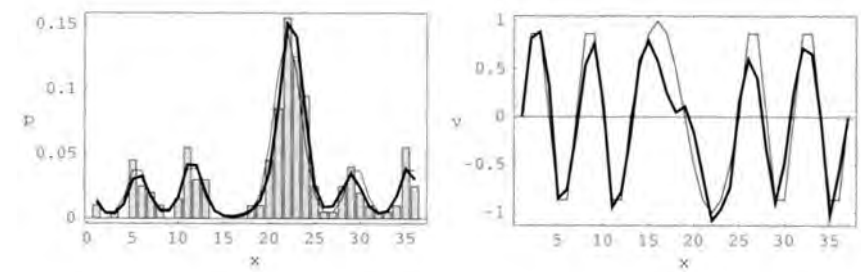


Figure 7.11: BIQS 11: Reconstruction of an approximately periodic potential with approximately periodic inverse covariance. Approximate periodicity is implemented by an inverse covariance $\mathbf{K}_0 = -\lambda(\Delta + \gamma\Delta_\theta)$ as in Eq. (7.68). (With $\gamma = 1.0$, $\lambda = 0.2$, a fixed $\theta = 6$, energy penalty term with $\mu = 1000$, and zero reference potential $v_0 \equiv 0$. Initial guess $v^{(0)} = v_0 \equiv 0$. All other parameters as for Fig. 7.8.)

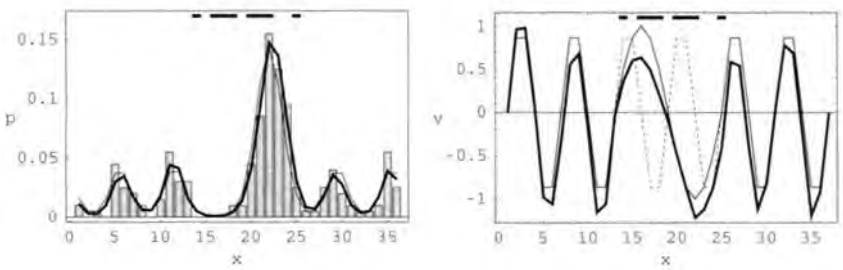


Figure 7.12: BIQS 12: Reconstruction of an approximately periodic potential with local switching 1. Shown is an example with local switching between a periodic and a zero reference potential. The black bars on top indicate regions where $B(x) = 1$, i.e., regions where impurities have been identified. (Prior of Eq. (7.69) with $\lambda_1 = 0.2$, $\lambda_2 = 0.2$, $\mu = 0$, and reference potential as in (7.66). The v–dependent function $B(x)$ was slowly changed from a sigmoid to a step function during iteration, keeping the threshold $\vartheta = 0.15$ fixed. All other parameters as in Fig. 7.8. Initial guess $v^{(0)}$ as for Fig. 7.10.)

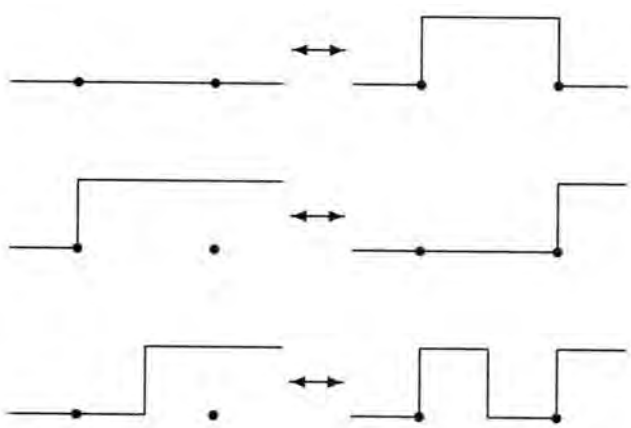


Figure 7.13: BIQS 13: Generation of new trial configurations for simulated annealing by selecting two points randomly and exchanging the values zero and one of the binary function in between. This mechanism has been used to optimize the binary functions $c(x)$ and $\theta(x)$.

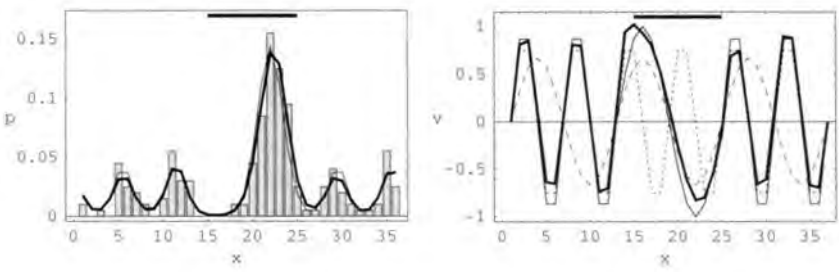


Figure 7.14: BIQS 14: Reconstruction of an approximately periodic potential with local switching 2. Shown is an example with local switching between two nonzero reference potentials. (Reference potentials v_1, v_2 given in (7.73) and (7.74). Prior of Eq. (7.71), with $\lambda_1 = \lambda_2 = 10$, $\mu = 0$. Step function for $B(x)$ with $\vartheta = 0$. An additional prior $p(B)$ on B has been included with $-\ln p(B)/10$ counting the number of discontinuities of the function $B(x)$. Other parameters as in Fig. 7.8.)

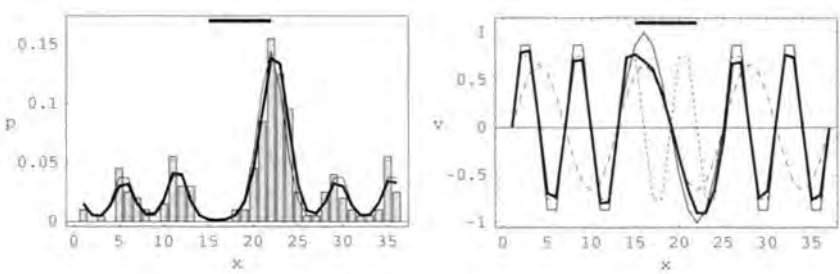


Figure 7.15: BIQS 15: Reconstruction of an approximately periodic potential with local hyperfield. (Prior of Eq. (7.75), with $\lambda_1 = 10$, $\lambda_2 = 1$, $\vartheta = 0$, $\mu = 0$, including a hyperprior $p(\theta)$ with $E_\theta/10$ counting the number of discontinuities of the hyperfield $\theta(x)$. Other parameters as in Fig. 7.8.)

7.2 Bayesian inverse time–dependent quantum theory (BITDQ)

7.2.1 Quantum time series

In this section we study the reconstruction of potentials by measuring particle positions in coordinate space for finite quantum systems in *time–dependent* states [298]. The presented '(Bayesian) inverse time-dependent quantum theory' (BITDQ) can easily be generalized to other forms of observational data.

To be specific, we study the following type of time series data generated by a quantum system: Preparing a particle in an eigenstate of the position operator with coordinates x_0 at time t_0, we let this state evolve in time according to the rules of quantum mechanics and measure its new position at time t_1, finding a value x_1. Continuing from this measured position x_1, we measure the particle position again at time t_2, and repeat this procedure until n data points x_i at times t_i have been collected. We thus end up with observational data of the form $D = \{(x_i, \Delta_i, x_{i-1}) | 1 \leq i \leq n\}$, where x_i is the result of the i-th coordinate measurement, $\Delta_i = t_i - t_{i-1}$ the time interval between two subsequent measurements, and x_{i-1} the coordinates of the previous observation (or preparation) at time t_{i-1}. In particular we consider systems with a time–independent Hamiltonian of the form $H = T + V$, consisting of a standard kinetic energy term T and a local potential $V(x, x') = \delta(x - x')v(x)$. In that case, our aim is the reconstruction of the function $v(x)$ from observational data D. The restriction to local potentials simplifies the numerical calculations. Nonlocal Hamiltonians can be reconstructed similarly. Again, being interested in predictions of future events, empirical learning requires *a priori* information to generalize from a finite number of observational data, i.e., to define the link between available and future data. It is the specific likelihood model, discussed next, which distinguishes BITDQ from other applications of the Bayesian approach.

The priors models for potentials discussed in the previous sections can be used for time–dependent problems as well. According to the axioms of quantum mechanics [503] the probability density that a particle is found at position x_i at time t_i, provided the particle has been at x_{i-1} at time t_{i-1}, is given by

$$p_i = p(x_i | \Delta_i, x_{i-1}, v) = |\psi_i(x_i)|^2, \tag{7.76}$$

(which can also be expressed in terms of a positive operator–valued measure [207,217]) where

$$\psi_i(x_i) = \langle x_i | \psi_i \rangle = \langle x_i | U_i x_{i-1} \rangle \tag{7.77}$$

are matrix elements of the time evolution operator

$$U_i = \exp(-i\Delta_i H), \tag{7.78}$$

setting $\hbar = 1$. The transition amplitudes (7.77) can be calculated by inserting orthonormalized eigenstates ϕ_α of H, with energies E_α,

$$\psi_i(x_i) = \sum_\alpha \exp\left(-i\Delta_i E_\alpha\right) \phi_\alpha(x_i)\phi_\alpha^*(x_{i-1}), \tag{7.79}$$

assuming H has a completely discrete spectrum, which can always be achieved by imposing zero or periodic boundary conditions.

Clearly it is straightforward to modify Eq. (7.76) for measuring observables different from the particle position. It is also interesting to note that the transition probabilities (densities) in Eq. (7.76) define a Markov process with $W_i(x \to x') = p(x'|\Delta_i, x, v)$. For real Hamiltonians with real boundary conditions (in coordinate representation) which have real eigenfunctions $\phi_\alpha(x)$ [336], the relation $W_i(x \to x') = W_i(x' \to x)$ holds. It follows that the detailed balance condition, $p_{\text{stat}}(x)W_i(x \to x') = p_{\text{stat}}(x')W_i(x' \to x)$, is fulfilled for a uniform $p_{\text{stat}}(x)$ which therefore represents the stationary state of the Markov process of repeated position measurements.

Having defined the likelihood model of BITDQ in Eq. (7.76), in the next step a prior for v has to be chosen, for example, a Gaussian prior process as in (7.50)

$$p_G(v) = \left(\det \frac{\mathbf{K}_0}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\left\langle v - v_0 \left| \mathbf{K}_0 \right| v - v_0 \right\rangle\right). \tag{7.80}$$

Typical examples are smoothness constraints on v which correspond to choosing differential operators for $\mathbf{K}_0$. Again, reference potentials can be made more flexible by working with potentials $v_0(x;\theta)$, parameterized by hyperparameters θ. In MAP approximation the optimal hyperparameters are determined by maximizing the posterior $p(v|D) \propto p(D|v)p(v)$ simultaneously with respect to θ and $v(x)$ [323]. A simplified procedure consists of using as the reference potential v_0 for a nonparametric reconstruction $v(x)$, a parametric approximation $v(\theta)$, which maximizes the likelihood $\prod_i p_i(\theta)$ (cf. Section 5.3).

If available, it is useful to include some information about the ground state energy $E_0(v)$ which helps to determine the depth of the potential. This can, for example, be a noisy measurement of the ground state energy which, assuming Gaussian noise, is implemented analogously to Eq. (7.33) by

$$p_E \propto \exp\left(-\frac{\mu}{2}\left(E_0(v) - \kappa\right)^2\right). \tag{7.81}$$

Combining (7.80) and (7.81) with (7.76) for n repeated coordinate measurements starting from initial position x_0, we obtain for the posterior

$$p(v|D) \propto p_G(v)\, p_E(v) \prod_{i=1}^{n} p_i. \tag{7.82}$$

To calculate the MAP solution $v^* = \text{argmax}_v p(v|D)$ we set the functional derivative of the posterior (7.82) or, technically more convenient, of its logarithm, with respect to v to zero,

$$0 = \delta_v \ln p(v|D) = \delta_v \ln p_G(v) + \delta_v \ln p_E(v) + \sum_i \delta_v \ln p_i, \tag{7.83}$$

where

$$\begin{aligned} \delta_v \ln p_G(v) &= -\mathbf{K}_0 (v - v_0), & (7.84)\\ \delta_v \ln p_E(v) &= -\mu\big(E_0(v) - \kappa\big)\, \delta_v E_0(v), & (7.85)\\ \delta_v \ln p_i &= 2\text{Re}[\psi_i^{-1}(x_i)\, \delta_v \psi_i(x_i)]. & (7.86) \end{aligned}$$

The functional derivative $\delta_v \psi_i$ can, according to Eq. (7.79), be obtained from $\delta_v \phi_\alpha$. As discussed in Section 7.1.1 the required $\delta_v \phi_\alpha$ and $\delta_v E_\alpha$ can be found by calculating the functional derivative of the eigenvalue equation $H\phi_\alpha = E_\alpha \phi_\alpha$

$$\begin{aligned} (E_\alpha - H)\,|\delta_v \phi_\alpha\rangle &= (\delta_v H - \delta_v E_\alpha)\,|\phi_\alpha\rangle\,, & (7.87)\\ \delta_{v(x)} E_\alpha &= \langle \phi_\alpha \,|\, \delta_{v(x)} H \,|\, \phi_\alpha \rangle = |\phi_\alpha(x)|^2. & (7.88) \end{aligned}$$

Using the Moore–Penrose inverse G_α of Eq. (7.15) we obtain as in Eq. (7.19), $\delta_v \phi_\alpha = G_\alpha (\delta_v H) \phi_\alpha$, or, written in components for normalized eigenfunctions,

$$\delta_{v(x)} \phi_\alpha(x') = \sum_{\gamma \neq \alpha} \frac{1}{E_\alpha - E_\gamma}\, \phi_\gamma(x') \phi_\gamma^*(x) \phi_\alpha(x). \tag{7.89}$$

Collecting the results it follows from Eq. (7.79),

$$\begin{aligned} \delta_{v(x)} \psi_i(x_i) &= \delta_{v(x)} \langle x_i | U_i\, x_{i-1} \rangle = \\ &\sum_\alpha \exp(-i\Delta_i E_\alpha) \Big[\big(-i\Delta_i |\phi_\alpha(x)|^2\big)\, \phi_\alpha(x_i) \phi_\alpha^*(x_{i-1}) \\ &+ \sum_{\gamma \neq \alpha} \frac{1}{E_\alpha - E_\gamma} \phi_\gamma(x_i) \phi_\gamma^*(x) \phi_\alpha(x) \phi_\alpha^*(x_{i-1}) \\ &+ \sum_{\gamma \neq \alpha} \frac{1}{E_\alpha - E_\gamma} \phi_\gamma^*(x_{i-1}) \phi_\gamma(x) \phi_\alpha^*(x) \phi_\alpha(x_i) \Big]. \end{aligned} \tag{7.90}$$

Inserting Eq. (7.88) for $\alpha = 0$ in Eq. (7.85) and Eq. (7.90) in Eq. (7.86), a MAP solution for the potential v will be found in the next section by iterating the stationarity equation (7.83) numerically on a lattice [40, 296, 303]. Clearly, such a straightforward discretization can only be expected to work for a low–dimensional x–variable. Higher dimensional systems usually require additional approximations, like the mean field approach discussed in Section 7.3.

7.2.2 Numerical case study: Time–dependent systems

As a next step we want to check the numerical feasibility of a nonparametric reconstruction of the potential v for a one–dimensional quantum system. For that purpose, we choose a system with the true potential

$$v_{\text{true}}(x) = \frac{c_1}{\sqrt{2\pi\sigma}} \exp \frac{-(x-c_2)^2}{2\sigma^2}, \tag{7.91}$$

where $c_1 = -10$, $c_2 = -2$, and $\sigma = 2$. An example of the time evolution of an unobserved particle in the potential v_{true} is shown in Fig. 7.16. As input for the reconstruction algorithm 50 data points x_i are sampled from the corresponding true likelihoods $p(x|\Delta_i, x_{i-1}, v_{\text{true}})$. A corresponding path of an observed particle is shown in Figs. 7.17 and 7.18.

It is interesting to remark that going to the limit of an infinite number of measurements, the time evolution of an ideal quantum system will finally stop (cf. Fig. 7.19). This is known as quantum Zeno effect [79, 135]. The quantum Zeno effect, however, is very difficult to observe [231] because real systems are typically destroyed (e.g., by relativistic particle production) if the frequency of the observations becomes too large. For example, it is in practice not possible to prevent the decay of a quantum state by frequent observation. In the contrary, it seems that a quantum Anti-Zeno effect (i.e., an increase of the decay rate with the measurement frequency) is easier to observe in real quantum systems [129, 268].

Besides a noisy energy measurement of the form (7.81) we include a Gaussian prior (7.80) with truncated Radial Basis Function inverse covariance as in Eq. (7.43),

$$\lambda \mathbf{K}_0 = \lambda \sum_{k=0}^{3} (-1)^k \frac{\sigma_0^{2k}}{k! 2^k} \left(\frac{\partial^2}{\partial x^2} \right)^k . \tag{7.92}$$

To obtain an adapted reference potential v_0 for the Gaussian prior, a parameterized potential of the form

$$v_0(a, b, c) = \min[0, a(x-b)^2 + c] \tag{7.93}$$

is optimized with respect to a, b, c by maximizing the 'extended likelihood' $\sum_i \ln p_i(v_0) + \ln p_E(v_0)$. Using a parametric reference potential corresponds to the prior expectation that the potential to be reconstructed has a form similar to Eq. (7.93) (with similarity defined by the inverse covariance). This exemplifies how the results of parametric reconstructions can be used as input for a more flexible nonparametric approach. Finally, the stationarity equation (7.83) is solved by iterating according to

$$\begin{aligned} v^{(r+1)} &= v^{(r)} + \eta \Big[v_0 - v^{(r)} + \\ &\qquad (\lambda \mathbf{K}_0)^{-1} \Big\{ 2 \sum_i^n \text{Re}[\delta_v \ln \psi_i(x_i)] + \delta_v (\ln p_G + \ln p_E) \Big\} \Big], \end{aligned} \tag{7.94}$$

with p_G and p_E as defined in Eq. (7.80) or Eq. (7.81), respectively, and optimizing the step width η by a line search algorithm. The resulting nonparametric BITDQ solution v_{BITDQ} (see Fig. 7.20) is a reasonable reconstruction of v_{true} and clearly better than the best parametric approximation $v_0(a,b,c)$. It is only the flat area near the right border where, due to missing and unrepresentative data, the reconstruction differs significantly from the true potential.

In Fig. 7.21 the transition probabilities (densities)

$$p_{\text{true}}(x) = \frac{1}{n}\sum_{i=1}^{n} p(x|\Delta_i, x_{i-1}, v_{\text{true}}) \tag{7.95}$$

under the true potential are compared with the corresponding empirical approximation $\widetilde{p}_{\text{emp}}(x) = \frac{1}{n}\sum_{i,j=1}^{n}\delta(x-x_i)$, as derived from the observational data D, and the reconstructed $p_{\text{BITDQ}} = \frac{1}{n}\sum_{i=1}^{n} p(x|\Delta_i, x_{i-1}, v_{\text{BITDQ}})$. Due to the summation over data points with different x_{i-1}, the quantities shown in Fig. 7.21 do not present the complete information which is available to the algorithm. Hence Fig. 7.22 depicts the corresponding quantities for a fixed x_{i-1}. In particular, Fig. 7.22 compares the reconstructed transition density (7.76) with the corresponding empirical and true transition densities for a particle having been at time t_{i-1} at position x_{i-1} =1. The BITDQ algorithm returns an approximation for all such transition densities.

Figs. 7.21 and 7.22 show that the reconstructed v_{BITDQ} tends to produce a better approximation of the empirical densities than the true potential v_{true}. Indeed, the error on the data or negative log–likelihood $\epsilon_D(v) = -\sum_i \ln p_i(v)$, being a canonical error measure in density estimation, is smaller for v_{BITDQ} than for v_{true}. A smaller λ, and thus a lower influence of the prior, produces a still smaller error $\epsilon_D(v_{\text{BITDQ}})$. At the same time, however, the reconstructed potential becomes more wiggly for smaller λ, being the symptom of the well known effect of overfitting. On the other hand, the (true) generalization error

$$\epsilon_g(v) = -\int dx\, dx' p(x) p(x'|x, v_{\text{true}}) \ln p(x'|x, v) \tag{7.96}$$

(with uniform $p(x)$) can never be smaller for the reconstructed v_{BITDQ} than for v_{true}. As is typical for most empirical learning problems, the generalization error $\epsilon_g(v_{\text{BITDQ}})$ shows a minimum as function of λ. It is this minimum which gives the optimal value for λ. Knowledge of the true model allows us to calculate in our case the generalization error exactly. If, as usual, the true model is not known, for example, cross–validation [12, 89, 169, 356, 456, 471, 472, 505–507, 509], bootstrap [118–120, 193] techniques, or the discrepancy method [67, 89, 122, 221, 222, 507] can be used to approximate the generalization error as a function of λ empirically. As discussed in Section 5.4.5 alternatively to optimizing λ or other hyperparameters one can integrate over them [323]. Similarly, studying the feasibility of a Bayesian Monte Carlo approach, contrasting the MAP approach, would certainly be interesting.

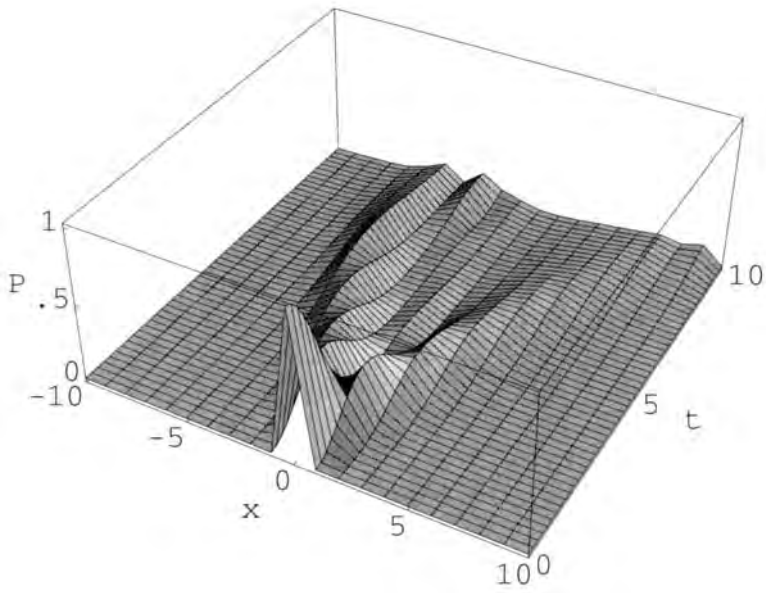


Figure 7.16: BITDQ 1: Time evolution of an unobserved particle with mass $m = 1$, started at time $t_0 = 0$ from $x_0 = 0$ in the potential (7.91). Shown is the transition density $p(x|\Delta t = t, x_0 = 0, v_{\text{true}})$.

In summary, in this section a Bayesian approach has been developed to solve the inverse problem for time–dependent quantum systems. The approach is able to handle quite general types of observational data and allows a flexible and explicit implementation of *a priori* information, the latter representing, for instance, data collected in the past or qualitative knowledge about the potential to be reconstructed. Classical examples of *a priori* information are smoothness and other approximate symmetries. Numerical results have been presented for a one dimensional model. The Bayesian approach to time–dependent inverse quantum theory is summarized in Fig. 7.23.

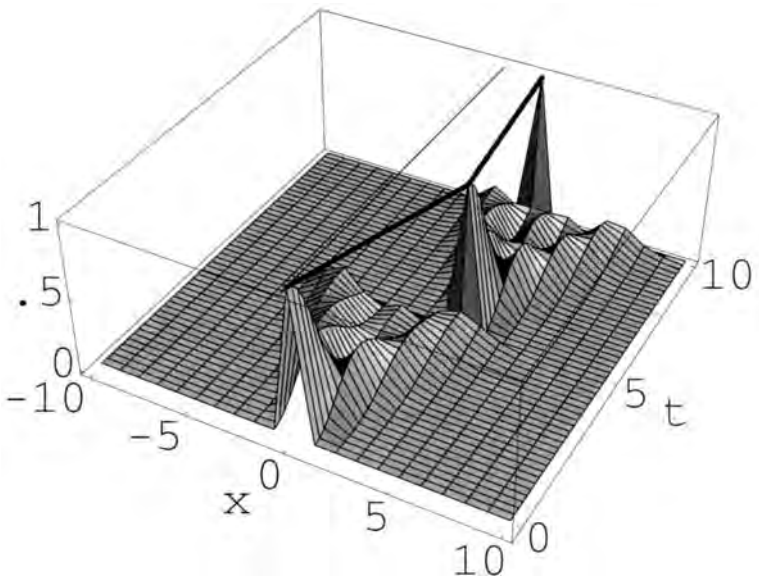


Figure 7.17: BITDQ 2: Time evolution of an observed particle with mass $m = 1$, started at time $t_0 = 0$ from $x_0 = 0$ in the potential (7.91) and observed again at $t_1 = 5$ and $t_2 = 10$.

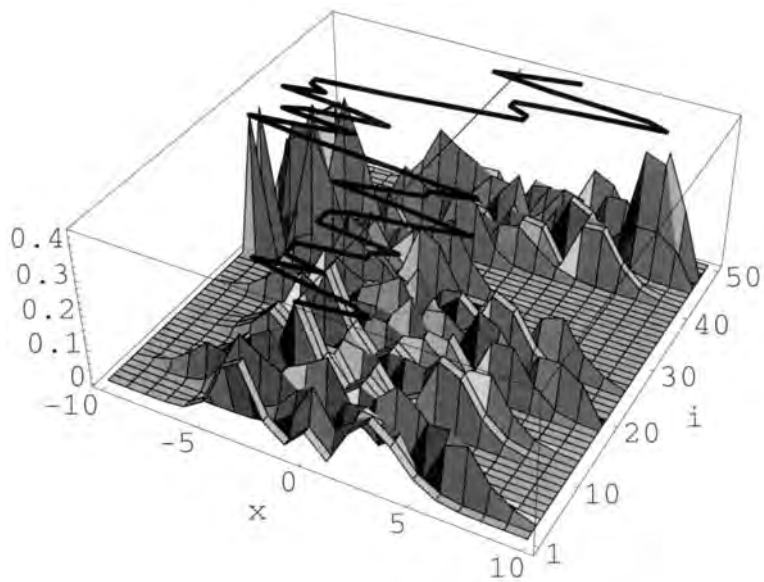


Figure 7.18: BITDQ 3: Time evolution of an observed particle with mass $m = 1$ in the potential (7.91). The figure shows for each data point i (measured at time t_i with time differences $\Delta_i = t_i - t_{i-1}$) the density $p(x|\Delta_i = 5, x_{i-1}, v_{\text{true}})$, starting from $x_0 = 0$. (Hence the density at $i = 1$ corresponds to that shown in Fig. 7.16 at $t = 5$.) The actual data points x_i have been sampled from these densities and form the observed path shown on top as a thick line.

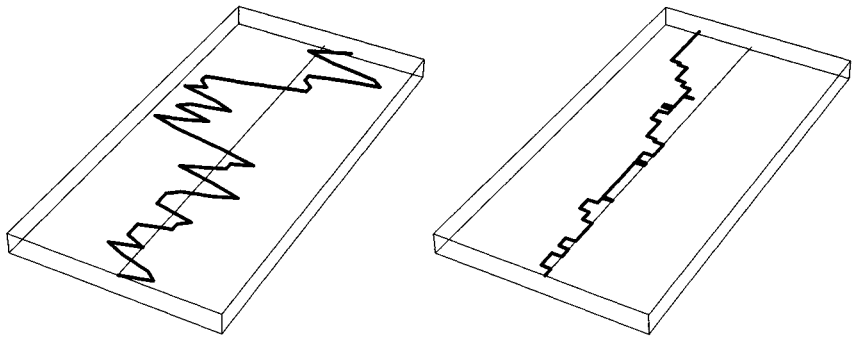


Figure 7.19: BITDQ 4: The quantum Zeno effect. The figure on the left hand side shows the path of a particle measured 50 times in the given time interval (cf. Fig. 7.18), the figure on the right hand side shows a sample path of an identical particle measured 1000 times in the same time interval. If the interval between consecutive measurements goes to zero, an ideal quantum system will not evolve in time anymore (quantum Zeno effect).

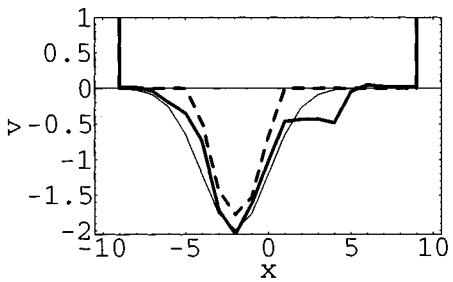


Figure 7.20: BITDQ 5: Numerical reconstruction of a potential from 50 coordinate measurements (see Fig. 7.18). Shown are the true potential v_{true} (thin line), the best parametric approximation used as reference potential v_0 (dashed line), and the reconstructed potential v_{BITDQ} (thick line). Parameters: $m = 1$, $\Delta_i = 5$, v_{true} of Eq. (7.91), v_0 of the form (7.93), Gaussian prior with $\lambda \mathbf{K}_0$ as in (7.92) with $\lambda = 0.1$ and $\sigma_0 = 3$, p_E of Eq. (7.81) with $\mu = 10$ and $\kappa = E_0(v_{\text{true}})$, periodic boundary conditions for ϕ_α, fixed boundary values $v(-10) = v(10) = 10^5$ for v_{BITDQ}, v_0 and v_{true}, calculated on a lattice with 21 points. Errors: $\epsilon_D(v_{\text{BITDQ}}) = 99.1$, $\epsilon_D(v_{\text{true}}) = 104.4$, $\epsilon_g(v_{\text{BITDQ}}) = 1.891$, $\epsilon_g(v_{\text{true}}) = 1.818$.

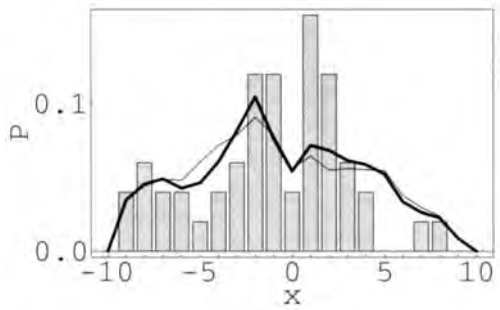


Figure 7.21: BITDQ 6: Empirical transition densities $\widetilde{p}_{\text{emp}}$ (bars), the corresponding true p_{true} (thick line) for v_{true}, and the reconstructed p_{BITDQ} (thin line).

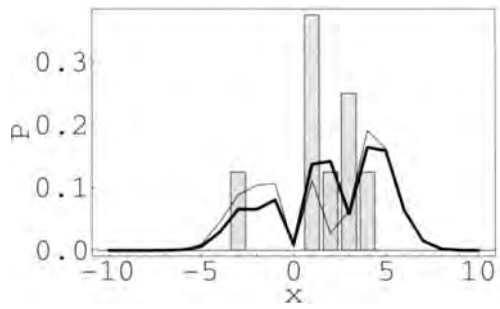


Figure 7.22: BITDQ 7: Conditional empirical transition densities. Same functions as in Fig. 7.21, but restricted to measurements of a particle which has been at position $x_{i-1} = 1$ at the time of the previous measurement.

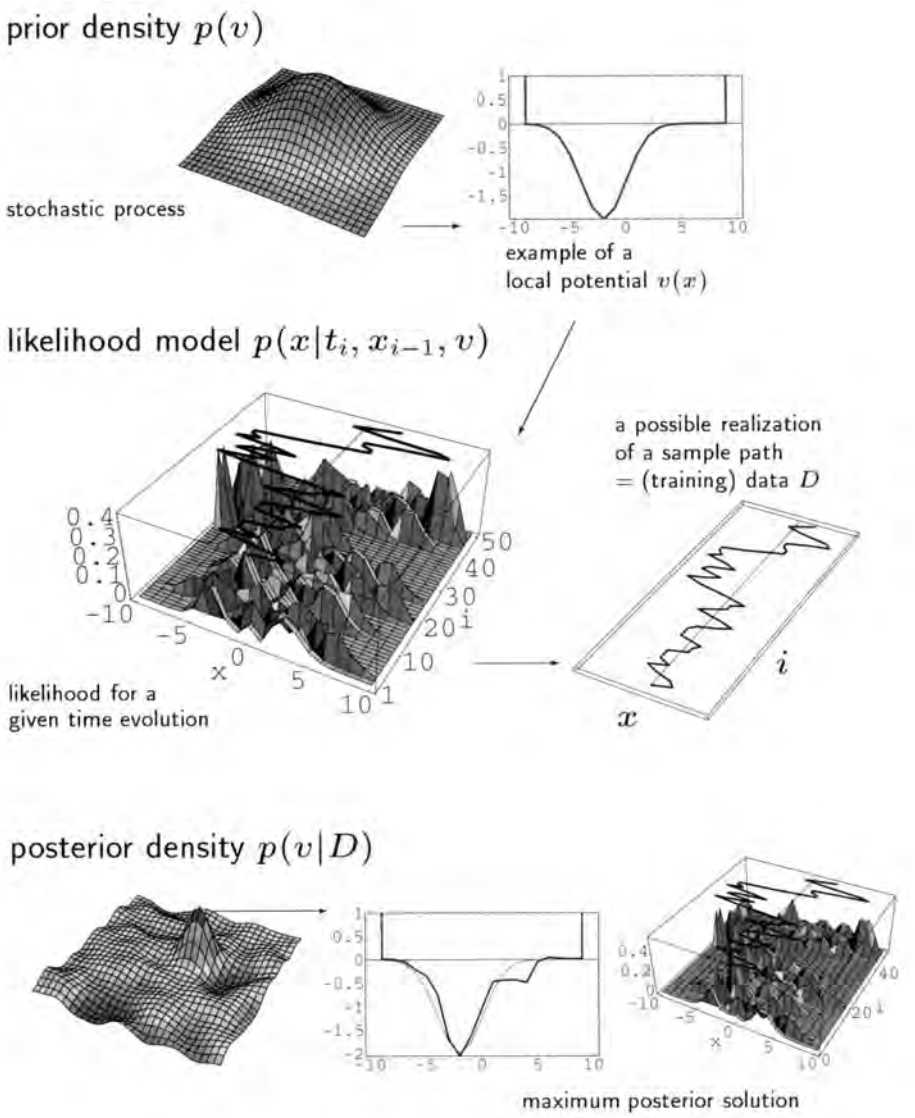


Figure 7.23: BITDQ 8: Scheme of Bayesian inverse time–dependent quantum theory. The 2–dimensional figures of the prior and posterior densities are symbolic representations of stochastic processes over potentials.

7.3 Bayesian inverse many–body theory

7.3.1 Systems of fermions

In this section the Bayesian approach for inverse problems will be applied to many–body systems [304]. To be specific, we will study the simultaneous measurement of the positions of N particles. We assume the measurement result to be given as a vector $\vec{x}_i$ consisting of N single particle coordinates $x_{i(j)}$. The treatment can easily be generalized to partial measurements of $\vec{x}_i$ by including an integration over components which have not been observed.

In particular, we will study two kinds of observational data $D = \{D_i | 1 \leq i \leq n\}$: (A) n simultaneous measurements of the coordinates of all N particles, (B) n measurements of the coordinates of a single particle. In case A, the ith measurement results in a vector $D_i = \vec{x}_i$, consisting of N components $x_{i(j)}$, each representing the coordinates of a single particle (which may also form a vector, e.g., a three dimensional one). According to quantum mechanics, the probability density of measuring the coordinate vector $\vec{x}_i$, given a potential v, becomes for a many–body canonical ensemble

$$p(\vec{x}_i|\hat{x}, v) = \mathrm{Tr}\Big(|x_{i(1)}, \cdots, x_{i(N)}\rangle \langle x_{i(1)}, \cdots, x_{i(N)}| \ \rho \Big), \tag{7.97}$$

which is a thermal expectation of squared many–body wave functions at $\vec{x}_i$

$$p(\vec{x}_i|\hat{x}, v) = \sum_\alpha p_\alpha |\psi_\alpha(x_{i(1)}, \cdots, x_{i(N)})|^2 = < |\psi(x_{i(1)}, \cdots, x_{i(N)})|^2 >, \tag{7.98}$$

with thermal probabilities $p_\alpha \propto \exp -\beta E_\alpha$ at inverse (physical) temperature β and many-body eigenfunctions ψ_α of a many–body Hamiltonian H with many–body energies E_α. For a system in the ground state ψ_0

$$p(\vec{x}_i|\hat{x}, v) = |\psi_0(\vec{x}_i)|^2 = |\psi_0(x_{i(1)}, \cdots, x_{i(N)})|^2. \tag{7.99}$$

In case B, the probability density for single particle coordinates x to take the values $D_i = x_i$ is obtained by summing in Eq. (7.98) over the unobserved coordinates

$$p(x_i|\hat{x}, v) = \sum_{x_{(2)}} \cdots \sum_{x_{(N)}} < |\psi(x_i, x_{(2)}, \cdots, x_{(N)})|^2 >, \tag{7.100}$$

in the ground state ψ_0,

$$p(x_i|\hat{x}, v) = \sum_{x_{(2)}} \cdots \sum_{x_{(N)}} |\psi_0(x_i, x_{(2)}, \cdots, x_{(N)})|^2. \tag{7.101}$$

Similarly, it is straightforward to obtain the likelihoods for measuring, for example, inter–particle distances, the particle number in a certain region, or particle momenta.

We will be interested in *fermions* for which the wave functions ψ_α and $|x_{i(1)}, \cdots, x_{i(N)}\rangle$ have to be antisymmetric. Considering a canonical ensemble, the density operator ρ has still the form of Eq. (7.3), but with H replaced by a many–body Hamiltonian. For fermions, it is convenient to express the many–body Hamiltonian in terms of creation and annihilation operators [56, 365]. A Hamiltonian with one–body kinetic energy $T = -(1/2m)\Delta$ and two–body potential V can then be written as

$$H = T + V = \sum_{ij} T_{ij}\, a_i^\dagger a_j + \frac{1}{4} \sum_{ijkl} V_{ijkl}\, a_i^\dagger a_j^\dagger a_l a_k, \tag{7.102}$$

with antisymmetrized matrix elements V_{ijkl}. Here $a_\alpha a_\gamma^\dagger + a_\gamma^\dagger a_\alpha = \langle \varphi_\alpha | \varphi_\gamma \rangle$ is equal to the overlap of the one–body orbitals $|\varphi_\alpha\rangle = a_\alpha^\dagger |0\rangle$ and $|\varphi_\gamma\rangle = a_\gamma^\dagger |0\rangle$ which are created or destroyed by the operators $a_\gamma^\dagger$ or a_α, respectively. Furthermore, $a_\alpha^\dagger a_\gamma^\dagger + a_\gamma^\dagger a_\alpha^\dagger = 0$, $a_\alpha a_\gamma + a_\gamma a_\alpha = 0$. A two–body eigenfunction ψ of the Hamiltonian (7.102), for example, can be expanded as

$$|\psi_\alpha\rangle = \sum_{\alpha,\gamma} c_{\alpha,\gamma}\, |\varphi_\alpha, \varphi_\gamma\rangle\,, \tag{7.103}$$

where $|\varphi_\alpha, \varphi_\gamma\rangle = a_\alpha^\dagger a_\gamma^\dagger |0\rangle$ denotes a Slater determinant being an antisymmetrized wavefunction.

The symmetrized version of a potential, local in relative coordinates, is

$$V_{x_{(1)}x_{(2)}x'_{(1)}x'_{(2)}} = v(z)\big[\delta_{x_{(1)},x'_{(1)}}\delta_{x_{(2)},x'_{(2)}} - \delta_{x_{(1)},x'_{(2)}}\delta_{x_{(2)},x'_{(1)}}\big], \tag{7.104}$$

where $z = |x_{(1)} - x_{(2)}|$ denotes the interparticle distance. Due to the antisymmetry the value of v at $z = 0$ does not enter the quantum mechanical formalism and one can, for example, choose $v(0) = 0$. (Alternatively, the value of $v(0)$ can be determined at a later stage of the calculation when prior information like smoothness of the function $v(z)$ is included.) Our aim is to reconstruct the function $v(z)$ (for $z > 0$) from observational data.

Having defined likelihoods $p(D|v)$ for many–body quantum systems in Eq. (7.98) and a prior $p(v)$ the next step of a Bayesian MAP approach is to solve the stationarity equation for the posterior $p(v|D)$

$$0 = \delta_{v(z)} \ln p(v) + \sum_i \delta_{v(z)} \ln p(\vec{x}_i|\hat{x}, v), \tag{7.105}$$

where $\delta_{v(z)} = \delta/\delta v(z)$ and $\ln p(v) = -\mathbf{K}_0(v - v_0)$ for a Gaussian process prior as in (7.50). An analogous equation has to be solved for the likelihood of case B (7.100). The MAP stationarity equation (7.105) is analogous to Eq. (7.34) of Section 7.1 with the likelihood terms adapted to the many–body case. Like in Eq. (7.34), a non–Gaussian energy prior p_U of the form of Eq. (7.33) can also be included in the many–body equation (7.105).

It is straightforward to show, by taking the derivative of the many–body Schrödinger equation $H\psi_\alpha = E_\alpha\psi_\alpha$ with respect to $v(z)$ that, similar to the one–body case,

$$\delta_{v(z)}E_\alpha = \frac{\langle\psi_\alpha|\delta_{v(z)}H|\psi_\alpha\rangle}{\langle\psi_\alpha|\psi_\alpha\rangle}, \tag{7.106}$$

$$|\delta_{v(z)}\psi_\alpha\rangle = \sum_{\substack{\gamma \\ E_\gamma \neq E_\alpha}} \frac{1}{E_\alpha - E_\gamma} |\psi_\gamma\rangle \langle\psi_\gamma| \delta_{v(z)}H |\psi_\alpha\rangle, \tag{7.107}$$

with a complete basis of eigenstates ψ_γ, energies E_γ, and fixing norm and phase by requiring

$$<\psi_\alpha|\delta_{v(z)}\psi_\alpha>= 0 \tag{7.108}$$

and, in the degenerate case $<\psi_\alpha|\delta_{v(z)}\psi_{\alpha'}> = 0$ for wave functions with equal energy. Furthermore, from $\delta_{v(z)}v(|x_{(1)} - x_{(2)}|) = \delta(z - |x_{(1)} - x_{(2)}|)$ follows directly (for one–dimensional x)

$$\delta_{v(z)}H = \frac{1}{2}\sum_x a_x^\dagger (a_{x-z}^\dagger a_{x-z} + a_{x+z}^\dagger a_{x+z})\, a_x. \tag{7.109}$$

From this the functional derivatives of the likelihoods, $\delta_{v(z)}p(x_i|\hat{x}, v)$, can be obtained. However, a direct numerical or analytical solution of the full inverse many–body equations is typically not feasible. To get a solvable problem we treat the many–body system in the next section in Hartree–Fock approximation.

7.3.2 Bayesian inverse Hartree–Fock theory (BIHF)

To tackle the inverse many–body problem we will refer to a Hartree–Fock approximation [56, 121, 365, 413]. For situations where a Hartree–Fock (HF) ansatz is not sufficient, the inverse problem has to be solved on top of other approximation schemes. A Random Phase Approximation or a full Time–Dependent Hartree–Fock Approximation, for example, would go beyond HF (see [166] and references above).

To this end we replace the full (hermitian) many–body Hamiltonian H by a (hermitian) one–body Hartree–Fock Hamiltonian $H^{HF} = \sum_{kl} h_{kl} a_k^\dagger a_l$ with matrix elements h defined, for example, in coordinate representation, as

$$h_{xx'} = T_{xx'} + \sum_k^N \langle x\varphi_k |V| x'\varphi_k\rangle, \tag{7.110}$$

the φ_k being the N–lowest (orthonormalized) eigenstates of h. The corresponding eigenvalue equation

$$h\varphi_k = \epsilon_k\varphi_k \tag{7.111}$$

is nonlinear, due to the φ_k–dependent definition (7.110) of h, and has to be solved by iteration. (The following calculation can be generalized to the case of non–hermitian H and h by distinguishing right eigenfunctions from left eigenfunctions, see [56,290].) The Hartree–Fock ground state is given by the Slater determinant $|\Phi_0\rangle = \det\{|\varphi_k\rangle\}$ made from the N–lowest orbitals, and has energy $E_0^{HF} = \sum_k^N t_{kk} + \frac{1}{2}\sum_{kl}^N v_{klkl} = \sum_k^N \epsilon_k - \frac{1}{2}\sum_{kl}^N v_{klkl}$. Considering the case of zero temperature, the many–body likelihood (7.99) in case A for the true ground state ψ_0,

$$p(\vec{x}_i|\hat{x},\rho(v)) = \langle\psi_0|\,\vec{x}_i\,\rangle\,\langle\vec{x}_i|\,\psi_0\,\rangle = |\psi_0(\vec{x}_i)|^2, \tag{7.112}$$

becomes in Hartree–Fock approximation

$$p_{\rm HF}(\vec{x}_i|\hat{x},\rho_{HF}(v)) = \langle\Phi_0|\,\vec{x}_i\,\rangle\,\langle\vec{x}_i|\,\Phi_0\,\rangle = |\Phi_0(\vec{x}_i)|^2 = |\det B_i|^2, \tag{7.113}$$

defining the overlap matrix B_i with matrix elements $B_{kl;i} = \langle x_{i(l)}|\,\varphi_k\,\rangle = \varphi_k(x_{i(l)})$. Analogously, the likelihood (7.101) in case B becomes

$$p_{\rm HF}(x_i|\hat{x},v) = \sum_{k=1}^{N} |\varphi_k(x_i)|^2. \tag{7.114}$$

To maximize the posterior, we have to calculate the functional derivative of the Hartree–Fock likelihood with respect to the potential. From (7.113) follows for case A $\delta_{v(z)}p_{\rm HF}(\vec{x}_i|\hat{x},v) = \Phi_0(\vec{x}_i)\delta_{v(z)}\Phi_0^*(\vec{x}_i) + \Phi_0^*(\vec{x}_i)\delta_{v(z)}\Phi_0(\vec{x}_i)$, so that expanding the determinant $\det B_i = \sum_l^N M_{kl;i}B_{kl;i}$ in terms of the cofactors, $M_{kl,i} = (B_i^{-1})_{lk}\det B_i$, yields

$$\langle\vec{x}_i|\,\delta_{v(x)}\Phi_0\,\rangle = \delta_{v(z)}\Phi_0(\vec{x}_i) = \sum_{kl}^{N} M_{kl;i}\,\delta_{v(z)}\varphi_k(x_{i(l)}). \tag{7.115}$$

Hence to obtain the functional derivatives with respect to $v(z)$ of the HF likelihoods, in case A given by Eq. (7.113), in case B by Eq. (7.114), the corresponding functional derivatives of the orbitals $\delta_{v(z)}\varphi_k$ are needed.

Again, we proceed by taking the derivative of Eq. (7.111) and obtain after standard manipulations as in Sections 7.1 and 7.2

$$\delta_{v(z)}\epsilon_k = \frac{\langle\varphi_k\,|\delta_{v(z)}h|\,\varphi_k\rangle}{\langle\varphi_k|\,\varphi_k\,\rangle}, \tag{7.116}$$

$$|\delta_{v(z)}\varphi_k> = \sum_{\substack{l \\ \epsilon_l\neq\epsilon_k}} \frac{1}{\epsilon_k-\epsilon_l}|\varphi_l><\varphi_l|\delta_{v(z)}h|\varphi_k>, \tag{7.117}$$

choosing $\langle\delta_{v(x)}\varphi_k|\,\varphi_l\,\rangle = 0$ for orbitals with equal energy. In contrast to Section 7.1, however, h and thus $\delta_{v(x)}h$ now obey a nonlinear equation. Indeed,

from Eq. (7.110) follows

$$\begin{aligned}\delta_{v(z)}h_{xx'} &= \sum_{j=1}^{N}\Big(\langle x\,\varphi_j\,|\delta_{v(z)}V|\,x'\,\varphi_j\rangle \\ &\quad + \langle x\,\delta_{v(z)}\varphi_j\,|V|\,x'\,\varphi_j\rangle \\ &\quad + \langle x\,\varphi_j\,|V|\,x'\,\delta_{v(z)}\varphi_j\rangle\Big). \qquad (7.118)\end{aligned}$$

Therefore, the derivative of the orbitals is found by inserting Eq. (7.118) into Eq. (7.117)

$$\begin{aligned}\delta_{v(z)}\varphi_k(x) &= \sum_{\substack{l\\ \epsilon_l\neq\epsilon_k}}\frac{\varphi_l(x)}{\epsilon_k-\epsilon_l}\sum_{j=1}^{N}\Big(\langle\varphi_l\varphi_j\,|\delta_{v(z)}V|\,\varphi_k\varphi_j\rangle \\ &\quad + \langle\varphi_l\delta_{v(z)}\varphi_j\,|V|\,\varphi_k\varphi_j\rangle \\ &\quad + \langle\varphi_l\varphi_j\,|V|\,\varphi_k\delta_{v(z)}\varphi_j\rangle\Big). \qquad (7.119)\end{aligned}$$

Recalling the definition of the antisymmetric matrix elements of V we finally arrive for one–dimensional x at the 'inverse Hartree-Fock equation'

$$\begin{aligned}\delta_{v(z)}\varphi_k(x) &= \sum_{\substack{l\\ \epsilon_l\neq\epsilon_k}}\frac{1}{\epsilon_k-\epsilon_l}\,\varphi_l(x) \\ &\times\sum_{j}^{N}\Bigg(\int dx'\,\varphi_l^*(x')\varphi_j^*(x'-z)\Big(\varphi_k(x')\varphi_j(x'-z)-\varphi_k(x'-z)\varphi_j(x')\Big) \\ &\quad+\int dx'\,\varphi_l^*(x')\varphi_j^*(x'+z)\Big(\varphi_k(x')\varphi_j(x'+z)-\varphi_k(x'+z)\varphi_j(x')\Big) \\ &\quad+\int dx'\,dx''\,\varphi_l^*(x')\Big(\delta_{v(z)}\varphi_j^*(x'')\Big)v(|x'-x''|) \\ &\qquad\times\Big(\varphi_k(x')\varphi_j(x'')-\varphi_k(x'')\varphi_j(x')\Big) \\ &\quad+\int dx'\,dx''\Big(\varphi_l^*(x')\varphi_j^*(x'')-\varphi_l^*(x'')\varphi_j^*(x')\Big) \\ &\qquad\times v(|x'-x''|)\varphi_k(x')\delta_{v(z)}\varphi_j(x'')\Bigg), \qquad (7.120)\end{aligned}$$

where for a Hamiltonian with real matrix elements in coordinate space the orbitals, and thus their functional derivatives, can be chosen real. Eq. (7.119) can be solved directly or, quite effectively, by iteration, starting, for example, with initial guess $\delta_{v(z)}\varphi_j(x)=0$. Because $\delta_{v(z)}\varphi_k(x)$ depends on z and x, Eq. (7.119), being the central equation of the Bayesian inverse Hartree–Fock approximation (BIHF), has the dimension of a two–body equation for

the lowest N orbitals. Introducing, analogous to B_i, the matrix $\Delta_i(z)$ with $\Delta_{kl;i}(z) = \delta_{v(z)}\varphi_k(x_{i(l)})$ (for $z > 0$, $1 \le l \le N$, $1 \le k \le N$, $1 \le i \le n$), it is straightforward to check that the derivatives of the logarithm of the HF likelihoods (7.113) and (7.114) are given by

$$\begin{aligned}\delta_{v(z)} \ln p_{\mathrm{HF}}(\vec{x}_i|\hat{x}, v) &= \frac{\sum_{kl}^{N} M_{kl;i}\Delta_{kl;i}(z)}{\det B_i} + \frac{\sum_{kl}^{N} M_{kl;i}^{\dagger}\Delta_{kl;i}^{\dagger}(z)}{\det B_i^{\dagger}} \\ &= 2\mathrm{Re}\left[\mathrm{Tr}(B_i^{-1}\Delta_i(z))\right], \end{aligned} \tag{7.121}$$

and

$$\delta_{v(z)} \ln p_{\mathrm{HF}}(x_i|\hat{x}, v) = 2\frac{\mathrm{Re}\left[\sum_{k=1}^{N} \varphi_k^*(x_i)\delta_{v(z)}\varphi_k(x_i)\right]}{\sum_{l=1}^{N} |\varphi_l(x_i)|^2}. \tag{7.122}$$

(The freedom to linearly rearrange orbitals for each data point i within the Slater determinants $\det\{|x_{i(l)}\rangle\} = \det\{|\tilde{x}_{i(l)}\rangle\}$, or within $\det\{|\varphi_k\rangle\}$, makes it possible to diagonalize the matrix of overlaps $\langle\varphi_k|\, x_{i(l)}\rangle$ in new orbitals $|\tilde{x}_{i(l)}\rangle$, which are then linear combinations of the $|x_{i(l)}\rangle$ [289,290,302].) The stationarity equation (7.105) for case A, and analogously the equation for case B, can now be solved by iteration like in (7.39)

$$v^{(r+1)} = v^{(r)} + \eta\left[v_0 - v^{(r)} + \mathbf{K}_0^{-1}\sum_i \delta_v \ln p_{\mathrm{HF}}(\vec{x}_i|\hat{x}, v^{(r)})\right], \tag{7.123}$$

choosing a positive step width η and starting from an initial guess $v^{(0)}$.

In conclusion, reconstructing a potential from data by BIHF is based on the definition of a prior for v and requires

1. for the actual guess $v^{(i)}$ the calculation of the single particle orbitals, defined in (7.111) as solutions of the direct (one–body) Hartree–Fock equation,

2. the corresponding derivatives of the likelihoods (7.121) or (7.122), respectively, obtained by solving the (two–body–like) equation (7.119),

3. needing as input for each iteration step (7.123) to solve the stationarity equation (like Eq. (7.105) in case A) for the potential. The procedure is iterated with the new guess $v^{(i+1)}$ until convergence is achieved.

7.3.3 Numerical case study: Inverse Hartree–Fock theory

To test the numerical feasibility of the BIHF we choose as Hamiltonian $H = T + V_1 + V$, consisting of a local two–body potential V to be reconstructed and a local one–body potential V_1, with diagonal elements $v_1(x) = 0$ for

$-a \leq x \leq a$ and $v_1(x) = \infty$ elsewhere, to confine the particles and to break translational symmetry.

To check the validity of the BIHF we must be able to sample artificial data from the exact true many–body likelihood (7.99) for a given true potential V_{true}. As this requires the corresponding many–body problem to be solved exactly, we have chosen a two–particle system with one–dimensional x (on a grid with 21 points, $a = 10$) for which the true ground state ψ_0 can be calculated by diagonalizing H numerically. We want to stress, however, that, in the case of real data, application of the BIHF to systems with $N > 2$ particles is straightforward and only requires to solve Eq. (7.119) for N orbitals instead for two orbitals.

We selected a local true two–body potential V_{true} with diagonal elements $v_{\text{true}}(z) = b/(1 + e^{-2\gamma(z-L/2)/L})$ and parameter values ($b = 1$, $\gamma = 10$, $L = 2a + 1 = 21$ for mass $m = 0.1$) for which the iteration of the HF equation (7.111) converges. (For two–body systems the HF iteration leads easily to oscillations.) Calculating the true ground state ψ_0 for V_{true} and the corresponding true likelihoods (7.99) and (7.101) we then sampled: (A) $n = 100$ two–particle data from the true likelihood (7.99) and (B) $n = 100$ single particle data from the true likelihood (7.101). (See Figs. 7.24a and 7.25a.)

The calculations have been done for a Gaussian prior process $p(v)$ as in (7.50) with $\lambda(\mathbf{I} - \Delta)/2$ (with identity $\mathbf{I}$, $\lambda = 5$) as inverse covariance $\mathbf{K}_0$, and a reference potential $v_0(z)$ of the form of v_{true}, but with $\gamma = 1$ (so v_0 becomes nearly linear in the shown interval). Furthermore, we have set all potentials to zero at the origin and constant beyond the right boundary. The reconstructed potential v_{BIHF} has then been obtained by iterating according to Eq. (7.123) and solving Eqs. (7.111) and (7.119) within each iteration step.

The resulting BIHF likelihoods $p(\vec{x}_i|\hat{x}, v_{\text{BIHF}})$ (case A, see Fig. 7.24a) or $p(x_i|\hat{x}, v_{\text{BIHF}})$ (case B, see Fig. 7.25a) indeed fit well the true likelihoods $p(\vec{x}_i|v_{\text{true}})$ or $p(x_i|v_{\text{true}})$, respectively. Fig. 7.24a shows instead of the two–dimensional $p(\vec{x}_i|v)$ for vectors $\vec{x}_i$ the one–dimensional

$$p(z|v) = \sum_x [p(x, x - z|v) + p(x, x + z|v)] \tag{7.124}$$

for the inter–particle distance z. In particular, in case A the reconstructed likelihood $p(z|\hat{x}, v_{\text{BIHF}})$ is over the whole range an improvement over the reference likelihood $p(z|\hat{x}, v_0)$, while in case B the BIHF solution $p(x_i|\hat{x}, v_{\text{BIHF}})$ is nearly exactly the same as the true likelihood $p(x_i|\hat{x}, v_{\text{true}})$. The perfect result for case B is due to the fact that reconstructing the likelihood $p(x_i|\hat{x}, v)$ for single particle data is a much simpler task than reconstructing the full $p(\vec{x}_i|\hat{x}, v)$.

The situation is more complex for potentials (Figs. 7.24b and 7.25b). Firstly, one sees that the correlation information contained in the two–particle data of case A yields a better reconstruction for v_{BIHF} than the less informative single particle data of case B. In both cases, however, the true potential

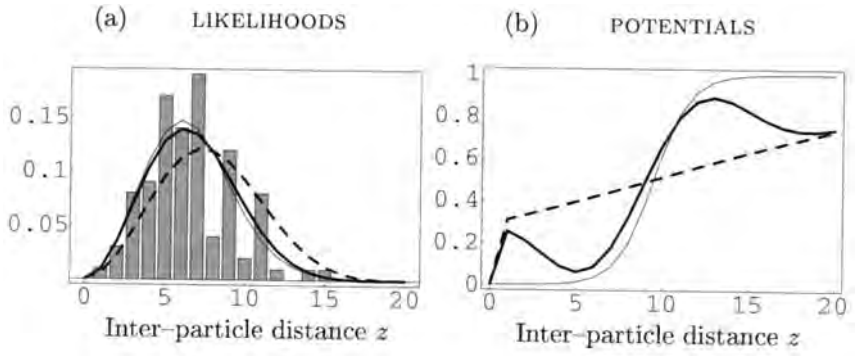


Figure 7.24: BIHF 1: Two–particle data (case A). (a) True likelihood $p(z|v_{\text{true}})$ (thin line), empirical density (relative frequency) $p_{\text{emp}}(z) = (1/n)\sum_{i=1}^{n}\delta_{z,|x_{i(1)}-x_{i(2)}|}$ (bars) obtained from the $n = 100$ sampled two–particle data $\vec{x}_i$, reference likelihood $p(z|v_0)$ (dashed), and reconstructed likelihood $p(z|\hat{x}, v_{\text{BIHF}})$ (thick line) as function of the inter–particle distance $z = |x_{(1)} - x_{(2)}|$. (b) True potential $v_{\text{true}}(z)$ (thin line), reference potential $v_0(z)$ (dashed), and reconstructed BIHF potential $v_{\text{BIHF}}(z)$ (thick line) as function of the inter–particle distance z.

is only well approximated at medium inter–particle distances. For large and small distances, on the other hand, the BIHF solution is still dominated by the reference potential of the prior $p(v)$. This effect is a consequence of the lack of empirical data in those regions: The probability density to find particles at large distances is small, because the true potential has its maximum at large distances. Also, measuring small distances is unlikely, because antisymmetry forbids two fermions to be at the same place. In such low data regions one must therefore rely on *a priori* information.

Fig. 7.26 summarizes the Bayesian inverse Hartree–Fock approach.

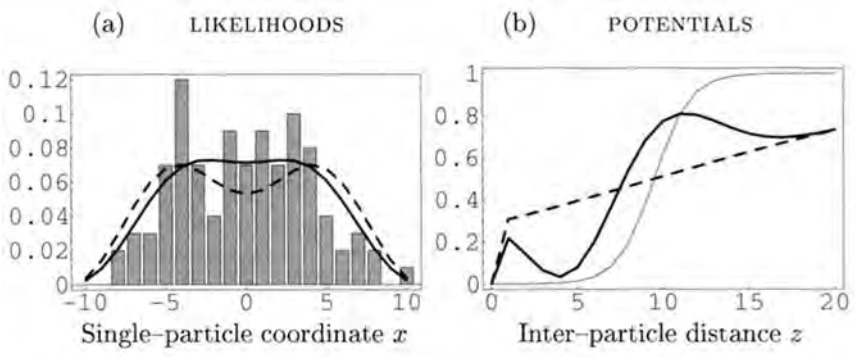


Figure 7.25: BIHF 2: Single particle data (case B): (a) True likelihood $p(x|v_{\text{true}})$ (occluded by thick line), empirical density $p_{\text{emp}}(x) = (1/n)\sum_{i=1}^{n}\delta_{x,x_i}$ (bars) obtained from the $n = 100$ sampled single particle data x_i, reference likelihood $p(x|v_0)$ (dashed), and reconstructed likelihood $p(x|\hat{x}, v_{\text{BIHF}})$ (thick line) as function of the single particle coordinate x. (b) True potential $v_{\text{true}}(z)$ (thin line), reference potential $v_0(z)$ (dashed), and reconstructed BIHF potential $v_{\text{BIHF}}(z)$ (thick line) as function of the inter-particle distance z.

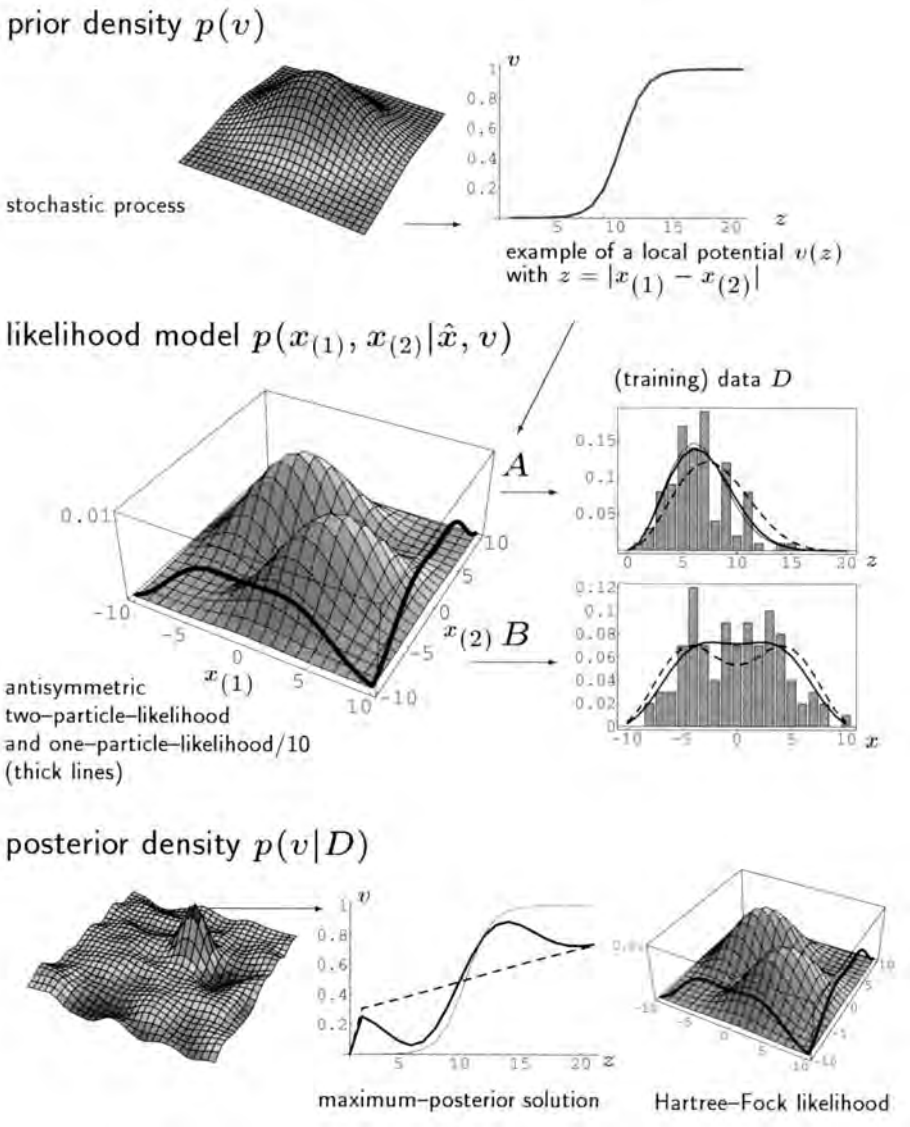


Figure 7.26: BIHF 3: Scheme of Bayesian inverse Hartree–Fock approximation. The 2–dimensional figures of the prior and posterior densities are symbolic representations of stochastic processes over potentials.

Chapter 8

Summary

In this book we wanted to develop a tool box for constructing prior models within a nonparametric Bayesian framework to empirical learning, and to exemplify its use for problems from different application areas. Nonparametric models, or field theories in the language of physics, allow typically a more explicit implementation of *a priori* information than parametric approaches. This is especially important for complex learning tasks which rely essentially on an adequate implementation of available *a priori* information.

After introducing the general nonparametric Bayesian framework (Chapter 2), the basic building blocks for constructing prior models, Gaussian process prior factors, were studied in Chapter 3 for general density estimation problems. In the formulation of the framework, it proved to be advantageous to include from the very beginning both, dependent variables ('y') and independent variables ('x'), as this allows density estimation, regression, and classification to be treated on the same footing. Already Gaussian prior factors, or generalized free fields in the language of physics, turned out to provide quite flexible prior models. By choosing the appropriate covariance they allow approximately invariant functions to be reconstructed, ranging from functions which are approximately invariant under infinitesimal translations (smooth functions), to approximately periodic functions and approximate fractals (Sections 3.6, 3.8.5, and 7.1.5). In contrast to exact symmetries, which are of fundamental importance for many theories in physics, the concept of approximate invariance or approximate symmetry seems to be more adequate for typical practical learning problems.

If a numerical solution by discretization becomes intractable for a nonparametric model, one can resort to a parametric model, for instance, additive models, decision trees, projection pursuit techniques, or neural networks. In Chapter 4 parametric methods are interpreted as variational approximations of nonparametric models. This defines a systematic procedure to derive prior terms for parametric approaches from an underlying nonparametric prior

model.

A large improvement in flexibility of prior models can be achieved by taking into account hyperparameters (Chapter 5). Technically the theory then becomes intrinsically non–Gaussian. In practice, the difficulty is to find the right balance between the additional flexibility introduced by hyperparameters and the information contained in the training data, because models which are too flexible do not generalize from training data to non–training data. For Gaussian prior factors hyperparameters parameterize the mean (Section 5.3) and/or the covariance (Section 5.4). For instance, parameterized means allow the solution of a parametric model to be taken as the mean of a Gaussian prior factor for a nonparametric model (Section 5.3.2). In image completion problems hyperparameters can adapt the scaling, rotational, and translational degrees of freedom of reference images (representing means of prior factors) to the training data (Section 6.6). Parameterizing the mean of a Gaussian process is especially convenient as this does not require adaption of a normalization factor.

Generalizing hyperparameters from numbers to functions leads to models with hyperfields, corresponding to full interacting field theories (Section 5.6). Hyperfields are useful to adapt prior models locally. An integer hyperfield, for instance, can reflect local distortions in an approximately periodic potential (Section 7.1.5). An example of a higher dimensional hyperfield is an optical flow field which adapts a reference image in an image completion problem to the training data. An alternative to models with hyperfields are the non–Gaussian models discussed in Section 5.7.

Models with an integer hyperparameter, where the summation over that hyperparameter can be carried out exactly, define the prior mixture models which are treated in Chapter 6. Mixtures of Gaussian prior factors are able to model multimodal prior surfaces and can thus take into account a moderate number of alternative Gaussian prior models simultaneously. Based on prior mixtures a method has been proposed to build nonparametric models from verbally formulated *a priori* knowledge.

To exemplify their use and to check computational feasibility, the basic variants of the prior models discussed in this book have been studied numerically for selected likelihood models, each representing a typical application class of empirical learning. The numerical case studies include: General density estimation with Gaussian prior factors (Section 3.4) and with prior mixtures (Section 6.3), Gaussian regression for approximately invariant functions (Section 3.8.5), and an image completion problem with hyperparameters and prior mixture model (Section 6.6).

As an example of special interest for physicists, Chapter 7 deals with the likelihood model of quantum theory. The empirical learning problem studied in this chapter is also called the inverse problem of quantum theory and consists of the reconstruction of potentials (or of Hamiltonians, density operators, respectively) from measuring a quantum system. For instance, such a

potential to be reconstructed can represent a surface with a periodic structure, distorted by impurities, located at unknown positions and of unknown form. While the development of nonparametric Bayesian methods for inverse quantum theory is still at its beginning, there are good reasons to expect that algorithms for solving inverse problems for quantum systems may become important already in the near future. Possible future application areas include atomic force microscopy, when scales are reached where a quantum mechanical description becomes mandatory, and quantum computing, where the result of a calculation has to be read off by measuring the final state of a quantum computer.

For inverse quantum theory the only genuine quantum mechanical input is the likelihood model as defined by the axioms of quantum theory. For the implementation of *a priori* information on potentials one can work with the same methods as developed for general density estimation problems. The reconstructed potential is found by maximizing the Bayesian posterior.

Specific advantages of this nonparametric Bayesian approach to inverse quantum theory are

- the possibility to combine heterogeneous data resulting from different quantum mechanical measurements, for example, momenta and coordinates,
- a flexible and explicit implementation of *a priori* information, and
- a straightforward adaption of the model to different learning scenarios.

Three scenarios have been discussed in detail:

- systems at finite temperature in Section 7.1,
- time–dependent systems in Section 7.2, and
- many–body systems in Section 7.3.

As a possible approximation scheme for many–body systems an inverse Hartree–Fock approach has been discussed. The nonparametric Bayesian approach turned out to be computationally feasible in the following numerical case studies:

- a problem of inverse quantum statistics in Section 7.1.3,
- the reconstruction of perturbed periodic potentials including the use of local hyperfields in Section 7.1.5,
- the reconstruction of potentials from a quantum mechanical time series in Section 7.2.2, and
- an application of inverse Hartree–Fock theory to a system of fermions in Section 7.3.3.

This book aimed to explore the flexibility of a nonparametric Bayesian approach. In conclusion, we may say that the principles of a nonparametric Bayesian approach can be adapted relatively easily to a large variety of different empirical learning situations. Numerical calculations, however, can easily become demanding. Nevertheless, simple one–dimensional or two–dimensional problems from general density estimation, regression, classification, and inverse problems in quantum theory can be solved directly in Maximum A Posteriori Approximation. For higher dimensional problems the Bayesian approach can be combined with other approximations which have been proven useful for the problem at hand, like a Hartree–Fock approximation in inverse many–body quantum theory. We may end with the remark, that in all applications of nonparametric Bayesian methods it is hardly possible to overemphasize the importance of *a priori* information.

Appendix A

A priori information and *a posteriori* control

In this appendix we discuss the role of *a priori* information from a strictly nonparametric point of view. From this nonparametric perspective, the influence of *a priori* information is much greater than for parametric models on which the discussion of *a priori* information is usually based. Indeed, for parametric models typically a large enough number of training data finally overwhelms the influence of *a priori* information (on the parameters). We have seen that this is not the case for nonparametric approaches where the number of training data points always remains small compared to the degrees of freedom of the model. (Clearly, the same is true for parametric models if not only *a priori* information on parameters is considered but also that part of the *a priori* information which corresponds to the selection of a specific parametric model.) The prominent role of *a priori* information in nonparametric models, however, seems very often to be underestimated. The crucial influence of *a priori* information is sometimes even seen as deficiency of the Bayesian approach (cf. Chapter 8 of [341]). Therefore this appendix discusses the role of *a priori* information from a strictly nonparametric point of view.

In Section 2.3 we have discussed three possible interpretations of *a priori* information, as stochastic control, empirical measurement, or structural implementation of dependencies. In all three interpretations, however, *a priori* information represents in practical applications mainly qualitative knowledge and does not arise from hard empirical data in the usual sense. The following excursion aims to show that a complete empirical verification of *a priori* information in a naive sense is hindered by principal difficulties in all three interpretations. First there is, especially from a nonparametric perspective, the problem that the number of introduced hidden variables and their possible relations can be very large or even infinite. In practice, however, there is

always a limit on the number of variables which can actually be controlled, measured, or structurally implemented. Second, there is the principal problem of controlling hidden variables. If there exists no possibility of a noisefree measurement of those variables (so they are hidden in all contexts), this leaves us with the problem of approximating a probability density indirectly. Third, there is the problem that hidden variables often represent variables which are not easily varied and are not random variables which can be sampled arbitrarily but are fixed, though unknown quantities. For example, being interested in finding an approximation for the radius of the earth from noisy measurements, it is difficult to think of a setting where several realizations of that radius can occur (see, for example, [314] and references therein). This problem of relating the mathematical concept of probability to single observed events has been extensively discussed and is at the core of the difference between Frequentist and Bayesian interpretations of statistical inference. Indeed, Frequentist statisticians have tried to base statistical inference strictly on the interpretation of probabilities as *asymptotical relative frequencies* [132, 133]. This, for example, does not allow for the assignment of probabilities to hypotheses (the hidden variables) and also leads to several principal difficulties, like the violation of the likelihood principle (see [33, 312] and Section 2.2.5). Furthermore, approximating probabilities by relative frequencies is typically only useful for low dimensional (parametric) problems and large numbers of available data. Bayesians have thus tried to base the interpretation of probability as *subjective degree of belief* on a consistent basis [87, 88, 102, 175, 234, 238, 311, 386, 446]. Interpreting probabilities as measure of plausibilities of propositions, Bayesians have often emphasized that probabilities are assigned, not measured. While this may be a suggestive formulation to characterize an important aspect of the Bayesian approach, it must be clearly said that such an assignment must have an empirical basis to allow generalization from training data to predict test data. Hence the problem of choosing appropriate *a priori* information is just a specific reformulation of the general induction problem, the problem of obtaining general laws from a finite number of empirical data. Nonparametric Bayesian approaches are an attempt to deal with the induction problem by formulating *a priori* information as explicitly as possible to facilitate its adaption to a specific learning problem under study. The success of many recent applications of empirical learning, like speech recognition or image analysis, relies essentially on appropriate problem–specific *a priori* information. Especially if only few training data are available, for example, because sampling is very costly, situation–specific *a priori* information is important.

In the previous sections we have seen that in a nonparametric setting it is only *a priori* information which enables generalization by providing the necessary link between measured training data and not yet measured (non–training) data. Referring to parametric approaches, however, the role of *a priori* information seems often to be underestimated. There are theorems, for example, proving that learning results become asymptotically independent of

a priori information on parameters if the number of training data goes to infinity. This, however, is possible only if the space of hypotheses ϕ is already sufficiently restricted[1] and if *a priori* information means knowledge in addition to that restriction. In particular, if there are x–values for which no training data are available, then learning for such x must refer to the mutual information of the likelihood at x and the available training data. Similarly, different y–values must be related by *a priori* information. Otherwise training would be useless for non–training situations. This also means that the generalization to non–training situations can be arbitrarily modified by varying the *a priori* information.

To make this point clear, consider the rather trivial situation of learning a deterministic function $\phi(x)$ where the x can take the two values x_1 and x_2. Having measured $\phi(x_1) = 5$, for example, then 'learning' $\phi(x_2)$ is not possible without linking it to $\phi(x_1)$. Such prior knowledge may have the form of a 'smoothness' constraint, say $|\phi(x_1) - \phi(x_2)| \leq 2$ which would allow a learning algorithm to 'generalize' from the training data and obtain $3 \leq \phi(x_2) \leq 7$. Obviously, arbitrary results can be obtained for $\phi(x_2)$ by changing the prior knowledge. This exemplifies that generalization can be considered as a mere reformulation of available information, i.e., of training data and prior knowledge. Except for such a rearrangement of knowledge, a learning algorithm does not add any new information to the problem. For a discussion of the related 'no–free-lunch' theorems see [528, 529].

Though extremely simple, this example nevertheless shows the principal problem. If the result of learning can be arbitrarily modified by changing *a priori* information, then it is absolutely critical which prior knowledge is implemented in a learning algorithm. This means in particular that successful learning requires correct prior knowledge, which therefore must have some empirical foundation just like standard training data have to be measured empirically. Otherwise the result of learning can not expected to be of any use.

In a nonparametric setting, however, empirical measurement of *a priori* information seems to be impossible. The reason is that we must link every possible test situation to the training data. We are not able to do this in practice if the number of potential test situations is larger than the number of measurements we are able to perform. Furthermore, it is not only a large number of hidden variables which can cause problems, certain variables may not be measurable at the time of training, for example, because they refer to the future or it is just too costly or impossible in practice. These arguments apply to a direct measurement of *a priori* information as well as to their control or structural implementation. In particular, formulating a general

[1] For example, a space of linear models with a fixed number of (linear) parameters would be sufficiently restricted. Models with one nonlinear parameter can already be too flexible to allow asymptotical convergence results. For instance, the parametric family $\sin(\xi x)$ with (nonlinear) parameter ξ has infinite Vapnik–Chervonenkis–dimension [494].

law one has, explicitly or implicitly, to refer to a 'ceteris paribus' condition, i.e., the constraint that all variables, not explicitly mentioned in the law, are held constant or are irrelevant. A 'ceteris paribus' condition represents (structural) *a priori* information, but an empirical verification would depend on all conceivable variables, also those not included in the model, and a clear definition of what 'being constant' means.

Consider again as an example the problem of learning a deterministic function $\phi(x)$. Then the *a priori* information of a bound valid for all $\phi(x_1) - \phi(x_i)$ requires all of those differences to be measured. Thus even if we take the deterministic structure of the problem for granted, the number of such differences is equal to the number of potential non–training situations x_i included in our model. Similarly, a bound on the function $\phi(x)$ would have to be verified for all x. Thus measuring *a priori* information does not require fewer measurements than directly measuring all potential non–training data.

Going to a probabilistic setting the problem of measuring *a priori* information gets even worse. Even if we restrict to Gaussian likelihoods with fixed variance, we require a prior for the regression function $\phi(x)$ for continuous x. If we choose a Gaussian process prior on $\phi(x)$ (see Section 3), this already requires the determination of a covariance and a mean function $t(x)$. But while already measuring a deterministic function $t(x)$ is in practice impossible for continuous x, in the probabilistic setting we have the additional complication that moments of a distribution can at most be measured asymptotically.

As a strictly empirical verification of *a priori* information therefore seems to be impossible, one might be tempted to rely on intuitive guesses to obtain the necessary prior for the specific situation under study. For example, some priors may appear more 'natural' than others. Indeed, the term *a priori* goes back to Kant [253], who assumed certain knowledge to be given necessarily *a priori* without reference to empirical verification. This means that one is either only able to produce correct prior assumptions for the relevant problems, because incorrect prior assumptions are 'unthinkable', for example, or that one is at least typically lucky in assigning the right prior process. But looking at the huge number of different prior assumptions which are usually possible (or 'thinkable') from a nonparametric point of view, there seems no reason why one should be lucky. In particular, because the result of learning in nonparametric approaches is almost completely determined by *a priori* information, one would have to expect that randomly chosen *a priori* information is almost never successful.

As an example consider again a Gaussian regression problem with a Gaussian process prior, which requires the choice of a prior mean function $t(x)$ and of a prior covariance. In that case the common choice of a zero mean function $t \equiv 0$ may seem to be a good first guess. But are there any arguments for using a zero (or constant) mean function? Clearly, one can always use a zero mean function for the prior process in Gaussian regression, at least formally, by subtracting a base line or reference function from ϕ. However, this for-

mal argument does not solve the problem, because it still leaves the base line or reference function to be determined, and without further information *any* function could be chosen as the reference function.[2]

An analogous example has been given in [238]: Having experimentally verified the gravitational law $s = a + vt + \frac{1}{2}gt^2$ for distance s, velocity v, and gravitational constant g at times $t_1, t_2, \ldots, t_n$ does not exclude laws like $s = a + vt + \frac{1}{2}gt^2 + f(t)(t - t_1)(t - t_2)\ldots(t - t_n)$ with any function $f(t)$ finite at times $t_1, t_2, \ldots, t_n$, meaning that the distance can be arbitrary if not observed. Similarly, like verification on training data alone does not allow prediction without further assumptions, also falsification of theories on the training data alone is useless if not related to future events by *a priori* information.

Trying to be cautious and using only weak, 'vague', or in some sense 'noninformative' priors (like Jeffreys' invariance prior which uses the Fisher information of the likelihood model under consideration [33,37,61,73,150,238, 308,424,457]) does also not solve the principal problem. Typically, the hope is that such priors, which may be, for example, a constant (improper) prior for a one–dimensional real variable, do not introduce a completely wrong bias, so that the result of learning is essentially determined by the training data. But besides the problem to define what exactly a noninformative prior has to be, such priors are in practice only useful if the set of possible hypotheses is already sufficiently restricted (to low dimensional parametric models), so 'the data can speak for themselves'. The problem remains to justify the priors which impose the necessary restrictions leading to a specific parametric model for which noninformative priors can be used. In contrast to parametric approaches, the many degrees of freedom in a nonparametric setting require in particular quite informative priors.

According to the arguments given so far, a complete empirical verification of *a priori* information seems impossible, a falsification useless, and finding correct *a priori* information by pure luck is very unlikely. Indeed, from a nonparametric point of view to learning, the primary problem is not the verification or falsification of theories, e.g., in the sense of Popper [341,406,407], but the question is how can it be possible to obtain a first at least reasonable theory or prior which is not falsified at once. (Model selection techniques are discussed in Chapter 5, formulated as adaption of hyperparameters.) Working in the extremely huge space of nonparametric models, which can never be completely determined by observed data alone, it looks as if successful learning is impossible. It is a simple fact, however, that learning can be successful. In that case there must be a way to measure or control (even nonparametric)

[2]The dependency of an empirical prediction on some underlying reference function has been exemplified in [235] as follows: Define a goy as a girl born before the year 2100 or a boy born thereafter. As up to now every goy has XX genotype using naive induction we would conclude that this will also be the case after the year 2100. But then a goy is defined to be a boy, which should have XY genotype if we believe in naive induction of our present observations of boys. (This example is known in several variations, for instance, with black and white swans replacing boys and girls.)

a priori information empirically.

The problem of determining nonparametric *a priori* information may be considered as artificial, arising from the introduction of a large number of *potential* test situations and a correspondingly large number of hidden variables ϕ representing what we call 'Nature'. In practice, however, the number of *actual* test data is always finite, just like the number of available training data is finite. This means that *a priori* information does have to link only the finite number of *actually* appearing test data to the training data, and not all *potentially* appearing test data. It is thus only a finite number of relations which must be under (at least approximate) control to allow successful generalization.[3]

The fact that only a finite number of dependencies is relevant opens the possibility to control *a priori* information empirically. However, because we do not know in general which test data will occur, such an empirical control can not take place at the time of training. Accordingly, *a priori* information has to be implemented at the time of measuring the test data [291,295].

Again a simple example may clarify this point. Consider the prior information that a regression function ϕ is bounded, $a \leq \phi(x) \leq b$, $\forall x$. A direct measurement of this prior assumption is practically not possible, as it would require every value $\phi(x)$ to be checked. An implementation within the measurement process is, however, trivial. One just has to use a measurement device which is only able to produce output in the range between a and b. This is a very realistic assumption and similar restrictions are valid for all real measurement devices. Values smaller than a and larger than b have to be filtered out or actively projected into that range. In case we nevertheless find a value out of that range we either have to adjust the bounds or we exchange the 'malfunctioning' measurement device with a proper one. Note that this range filter is only needed at the finite number of actual measurements. This gives an example of how *a priori* information can be implemented by *a posteriori control* at the time of testing.

Furthermore, we have discussed in Section 2.3.2 that a realistic measurement device does not only produce bounded output but also shows output noise and input noise, and often its measurement result is an average over many approximately independent variables. Such noise or averaging processes are usually understood to be part of the measurement process. Hence these processes are supposed to occur also for measuring test data and can thus be interpreted as *a posteriori* control. For example, a device measuring at $x + \Delta$ rather than at x, with Δ being a Gaussian random variable with zero mean, shows translational input noise. In Section 2.3.2 it has been shown that such devices with bounded output and translational input noise produce smooth functions [53, 286, 291, 514]. This may be an explanation for the success of

[3]This is related to Vapnik's distinction between induction and transduction problems [494]: In induction problems one tries to infer a whole function, in transduction problems one is only interested in predictions for a few specific test situations.

smoothness priors in practice.

As already discussed such additional noise or averaging processes (called structural priors in Section 2.3.1) still refer to a possibly large or infinite number of hidden variables and hence can not be completely verified empirically. Thus to be able to verify empirically a specific *a posteriori* control mechanism, it must depend only on visible variables. Those are the training data D, possibly a finite number of additional visible prior data D_0, and the new test situation x. Hence what has to be controlled is the predictive density $p(y|x, D, D_0)$ without reference to hidden variables. Thus, a learning algorithm predicting test data from training data has to be an equivalent of the real mechanisms which produce the test data. (In this sense parametric methods are adequate if the real mechanism which produces test data is known to be of parametric form with unknown parameters.) From this point of view *a priori* information corresponds to an explicit stochastic (*a posteriori*) control of y given the test situation and training data. In other words, the measurement process of test data must ensure that only test data (x, y) can be observed which fulfill all required dependencies. For example, if we require for a deterministic function $\phi(x)$ the *a priori* information $|\phi(x) - \phi(x + 1)| \leq 1$, then we have to use a measurement device which, for given training data D stored in some memory, only produces output consistent with that requirement.

Still this is not the end of the story, because probability densities, like the predictive density or probabilistic restrictions, are not observable quantities. Even for binary y one can only approximate a probability $p(y)$ by a relative empirical frequency in the ideal case where sampling a large enough number of observations under stationary conditions is possible. Probabilistic models can in practice at best asymptotically, but never strictly, be verified by empirical sampling. Hence, if only a few measurements are available for a given condition and an asymptotic interpretation is not possible, one has to refer to a different interpretation of probability. Indeed, from a Bayesian point of view probabilities represent degrees of belief [33, 238, 314]. In this interpretation the degree of belief can be based on known frequency but is not identical with frequency. Much work has been devoted to put the measurement of beliefs on a consistent decision theoretical basis [235–237, 277, 316, 411, 436, 504]. (Practical difficulties of subjective probability elicitation arise, for instance, from the fact that real decisions are typically not rational decisions [156, 248–250, 288, 454, 487, 488].) Beliefs in form of qualitative knowledge (which according to the previous discussion should be related to implicit or explicit *a posteriori* control) are thus required to determine prior distributions for which no physical measurement device exists. In practice, qualitative knowledge can be incorporated in quantitative models, for example, by using fuzzy–logic like techniques [265, 266, 291–295, 510]. Similarly, Section 3 discusses how to use Gaussian prior processes to approximate invariances, while Sections 5 and 6 show how for complex situations specific prior processes can be constructed from simple Gaussian building blocks using mixtures, hyperparameters, and

hyperfields. Due to their flexibility, implementing problem–specific *a priori* information is crucial for nonparametric approaches.

The interpretation of *a priori* information as *a posteriori* control is related to the point of view of constructivism, which is relatively popular in contemporary philosophy (see, for example, [307,513]). It has been characterized by Vico (1710) in the sentence: *Verum ipsum factum* — the truth is the same as the made [498]. 'Construction of reality' does here not mean that arbitrary laws of nature can be constructed by 'free will', but it emphasizes that a coupling between (the model used by) the observer and the observed system is necessary to be able to learn from data.[4,5]

Hence from a constructivistic point of view the only possibility of a strict empirical verification of *a priori* information consists of an explicit control of the test data and their required dependencies to the training data.[6] In this interpretation empirical verification of *a priori* information corresponds to a more or less tautological implementation of all required dependencies. In practice, however, working with hidden variables often turns out to be very useful. For example, it can be easier to work with a posterior density instead of memorizing all training data. One should keep in mind, however, that probabilistic models, especially if they contain hidden variables, can not be strictly empirically verified. Useful priors can also not result from a pure random trial and error strategy. The reason is that the number of possible prior distributions is too large and wrong priors usually lead to completely wrong generalization. This appears especially clear in a nonparametric formulation of learning. Thus even if not formulated explicitly, the validity of

[4]Constructivism has recently been applied mainly to psychology and sociology (see [36, 188, 439, 502, 512, 513] and references therein). A detailed study of 'intersubjective' construction processes for natural sciences from a practical point of view is, to the best of my knowledge, still lacking. (A more philosophically oriented 'constructive empiricism' has been developed, for example, in [490], discussed in Essay 5 of [74].) One step in this direction is provided by the approach of constructive mathematics which works only with mathematical objects which can be constructed explicitly [54]. The role of active control and the impossibility to measure all necessary prior information completely is also often emphasized by statisticians and philosophers which prefer a 'causal calculus' over a 'probability calculus' [116,387,389]. For instance, J. Pearl ('The bulk of human knowledge is organized around causal, not probabilistic relationships, and the grammar of probability calculus is insufficient for capturing those relationships' [390]) and N. Cartwright ('no causes in, no causes out' [75]) state that causal assumptions are not testable in observational studies.

[5]If control strategies are available it can be very useful to include those explicitly in the probabilistic model. (An example from economics is provided by the celebrated Black–Scholes formula for option pricing where the price of an option is determined by reference to an optimal trading strategy (Delta–hedging). Under the specific assumptions of the Black–Scholes model this eliminates all random effects and makes the option price predictable [105, 228, 287, 523].)

[6]The 'classical' Bayesian view of Jeffreys, choosing as a starting point beliefs and not the dependencies between training and test data, 'rejects the attempt to reduce induction to deduction' as attributed to the Frequentist approach (see preface of [238]). In a constructivistic interpretation of learning, prediction appears again as deduction based, however, on training data and *a posteriori* control.

probabilistic models has to be based on some implicit processes of *a posteriori* control which is reflected in the qualitative knowledge used to build the model. Otherwise learning can not be successful.

Summarizing we may say, that from a constructivistic point of view implemented *a priori* information should provide a model for the implicit training data–dependent processes of *a posteriori* control at the time of testing the theory, contrasting the classical point of view of *a priori* information as information about a hidden state of Nature. As, however, the actual processes of *a posteriori* control are typically also not known precisely, the two interpretations do not necessarily differ in practical applications of the Bayesian formalism. In both cases *a priori* information is obtained from practical experience with similar tasks. Such expert knowledge is typically of a more qualitative than quantitative nature. Techniques to translate such qualitative knowledge into a quantitative model are available and have been discussed in more detail in Chapters 5 and 6. For practical model building all of the discussed interpretations of *a priori* information, e.g., in terms of controlled, empirically measured, or structurally implemented dependencies, can be useful. In particular, the constructivistic point of view reminds us that it can be worthy to analyze the relevant measurement process for which learning is intended.

This page intentionally left blank

Appendix B

Probability, free energy, energy, information, entropy, and temperature

B.1 Statistics and statistical mechanics

Both disciplines, statistics and statistical mechanics, deal with probabilistic models. Their differences in language and methods can be traced back to differences in their typical applications.

Statistical mechanics has been developed for extremely large systems, like they appear in condensed matter physics. Typical systems of statistical mechanics are of high regularity, defined on a two or three dimensional grid with local variables having the same range of possible values (e.g., ± 1 for spins or real numbers for scalar fields) and only local interactions. There are relatively few prototypical systems, which have been studied extensively, for example, the celebrated Ising model. A complex machinery has been developed to obtain results with very high accuracy but requiring long and costly calculations.

Statistics, on the other hand, is mainly interested in the solution of a larger variety and more application oriented problems. Compared to prototypical systems studied in statistical mechanics, the corresponding models are therefore often smaller but less regular. The main practical problem consists in constructing adequate models. The models are usually not expected to allow very precise predictions. Hence, there is no need to achieve, by costly calculations, a numerical accuracy which is beyond the validity of the model. Needed are fast, flexible, and robust algorithms.

However, due to the increasing computing power becoming widely available, the gap between the two disciplines gets smaller. There is now a growing

and seminal interaction between the two disciplines, like in the development of Monte Carlo methods or graphical models. Typical areas where methods of statistical mechanics can be easily applied have a large amount of quantitative data of the same kind, organized on a one, two, or three dimensional grid with dependencies dominated by local interactions. This, for example, is the case in Bayesian image reconstruction or when predicting financial time series. As it therefore becomes important to understand the languages of both approaches, we will discuss in the following the relations between concepts of statistics and statistical mechanics.

We begin with some remarks concerning the construction of probability distributions:

1. (Normalization constants or partition sums.) The specification of a probability $p(x)$ starts with unnormalized numbers $Z(x)$. To ensure the normalization condition $\int dx\, p(x) = 1$ a normalization constant $Z = \int dx\, Z(x)$ has to be calculated. For simple systems like a dice this is easy. For large systems, like in typical systems studied in statistical mechanics, this can be a highly non–trivial task. Thus, unnormalized functions $Z(x)$ instead of probabilities $p(x)$ are the natural starting point for large systems. Despite the fact that the Z are not probabilities because they are not normalized, we will call them in the following *'unnormalized probabilities'* or *partition sums.*

2. (Log–probabilities, information and energy.) A large system is usually constructed out of simpler subsystems, with the probabilities of the subsystems combined by multiplication, according to $p(x_1, x_2) = p(x_1)p(x_2|x_1)$. Sums, however, are easier to handle than products and represent a somewhat more intuitive concept. So it is easier to deal with infinite sums, or in the continuous case, with integrals, than with infinite or continuous products. This means that large systems are easier constructed in terms of *logarithms of probability* $\ln p(x)$. We will show in the following how *information* is related to expectations of the logarithms of probabilities and *energy* to expectations of the logarithm of unnormalized probabilities Z.

3. (Conditional probabilities and disordered systems.) Instead of directly constructing a complicated probability distribution $p(x)$ it is often easier to break $p(x)$ down in simpler parts. This is done by selecting conditions y under which $p(x|y)$ and $p(y)$ are relatively easy to specify. The total probability is then obtained by combining the alternatives y according to $p(x) = \int dy\, p(y)p(x|y)$. For example, it can be easier to specify probabilities of symptoms for given specific diseases and probabilities of diseases than to write down directly a probability for the symptoms averaged over all diseases in one step. This, however, also means that due to the normalization requirements $\int dx\, P(x|y) = 1$, a normalization

constant $Z(y)$ over x must be calculated for every y. In statistical mechanics such models are called *disordered systems*. Such averages over energy functions occur, for example, for spin glasses [128,338,380].

Thus, while statistics can often be formulated directly in terms of probabilities, statistical mechanics uses a formulation in terms of logarithms of unnormalized probabilities. In the following their relations to the concepts of energy and free energy will be discussed.

B.2 Probability

Especially for large and complex systems, it is convenient to work with the logarithm of 'unnormalized probabilities' instead of directly with probabilities.

B.2.1 Normalization factors or partition sums

Let X be a random variable with possible values $x \in \mathcal{X}$ and assume a probability measure $p(A)$ defined on a σ–algebra of possible events $\mathcal{A}$ being subsets of a set $\mathcal{X}$. The event x may be represented by a vector of real numbers.

Denoting unnormalized probabilities by $Z(x) \propto p(x)$ we write

$$p(x) = \frac{Z(x)}{Z(\mathcal{X})} \tag{B.1}$$

and for general events A with $A \subseteq \mathcal{X}$, i.e., including $A = \mathcal{X}$

$$Z(A) = \int_{x \in A} dx\, Z(x) \propto p(A) = \int_{x \in A} dx\, p(x), \quad p(A) = \frac{Z(A)}{Z(\mathcal{X})}. \tag{B.2}$$

Introducing a second random variable Y and using the compatible normalization

$$Z(y) = Z(\mathcal{X}, y) = \int dx\, Z(x, y) \propto p(y) = \int dx\, p(x, y) \tag{B.3}$$

gives

$$Z(\mathcal{X}, \mathcal{Y}) = \int dx \int dy\, z(x, y) = \int dy\, z(y) = Z(\mathcal{Y}) \tag{B.4}$$

and therefore

$$p(y) = \frac{Z(y)}{Z(\mathcal{Y})} = \frac{Z(y)}{Z(\mathcal{X}, \mathcal{Y})} \quad \text{and } p(x|y) = \frac{p(x, y)}{p(y)} = \frac{Z(x, y)}{Z(y)}. \tag{B.5}$$

Choosing $Z(y)$ to be not equal but proportional to $Z(\mathcal{X}, y)$ one obtains

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{Z(x, y)\, Z(\mathcal{Y})}{Z(y)\, Z(\mathcal{X}, \mathcal{Y})}. \tag{B.6}$$

In slight generalization of the language of statistical physics we also call unnormalized probabilities Z *partition sums*.

B.2.2 Log–probabilities, bit numbers, and free energies

Log–probabilities L are defined by

$$L(x) = \ln p(x), \quad p(x) = e^{L(x)} \tag{B.7}$$

and for $A \subseteq \mathcal{X}$

$$L(A) = \ln p(A), \quad p(A) = e^{L(A)}. \tag{B.8}$$

In terms of log–probabilities L a product like

$$p(x, y) = p(x)p(y|x) \tag{B.9}$$

becomes a sum

$$e^{L(x,y)} = e^{L(x)+L(x|y)}. \tag{B.10}$$

Log–probabilities are widely used in practice because it is often more convenient to deal with sums than with products. Also the 'quenching' effect of the logarithm can be important in numerical calculations when $p(x)$ varies over several orders of magnitudes.

Especially common are negative log–probabilities b, also called *bit numbers*,

$$b(x) = -\ln p(x), \quad p(x) = e^{-b(x)} \tag{B.11}$$

and for $A \subseteq \mathcal{X}$

$$b(A) = -\ln p(A), \quad p(A) = e^{-b(A)}. \tag{B.12}$$

Analogously the *free energy* is defined for unnormalized probabilities Z by

$$F(x) = -\frac{1}{\beta} \ln Z(x), \quad Z(x) = Z(x|\beta) = e^{-\beta F(x)}, \quad p(x) = e^{-\beta(F(x)-F(\mathcal{X}))}. \tag{B.13}$$

and for $A \subseteq \mathcal{X}$

$$F(A) = -\frac{1}{\beta} \ln Z(A), \quad Z(A) = Z(A|\beta) = e^{-\beta F(A)}, \quad p(A) = e^{-\beta(F(A)-F(\mathcal{X}))}. \tag{B.14}$$

The dependency on β will not always be denoted explicitly in the following. For $x \in \mathcal{X}$ the factor $e^{-\beta F(x)}$ is also known as Boltzmann weight of x. The reasons for the introduction of the parameter β will be discussed later in detail.

B.3 Random variables

Information and energy are averages of special random variables. We discuss now their connection to bit numbers and free energy.

B.3.1 Averages

Recall the definition of an expectation or average of a random function $g(x)$ over $\mathcal{X}$

$$g_X = \langle g(x) \rangle_{\mathcal{X}} = \int dx\, p(x) g(x). \tag{B.15}$$

Analogously, we define an expectation over a subset $A \subseteq \mathcal{X}$

$$g_X(A) = \langle g(x, A) \rangle_A = \int dx\, p(x|A) g(x, A) = \int_{x \in A} dx\, \frac{p(x)}{p(A)} g(x, A) \tag{B.16}$$

using $p(x, A) = p(x)$ and $p(x|A) = p(x)/p(A)$ for $x \in A$. Using a second random variable Y this can be generalized to a conditional expectation of a random function $g(x, y)$ over event $Y = y$

$$g_X(y) = \langle g(x, y) \rangle_{\mathcal{X}|y} = \int dx\, p(x|y) g(x, y). \tag{B.17}$$

An average of form (B.16) is obtained by choosing in (B.17) a random variable Y taking the value y everywhere on A but not on its complement.

B.3.2 Information

A random variable C on $\mathcal{X}$ corresponds to a function $C(x)$ defined for every $x \in \mathcal{X}$. A special example is a transformation T of $p(x)$ which defines a corresponding random variable by $C(x) = T(p(x))$ for $x \in \mathcal{X}$. We will call in the following C the canonical random variable of the transformation T on $\mathcal{X}$. Specifically, for every distribution $p(x)$ the corresponding bit number $b(x) = -\ln p(x)$ can be considered as random function with the property of being defined not only x– but also $p(x)$–dependent. The canonical random variable on $\mathcal{X}$ for bit numbers, i.e., for the transformation $-\ln p(x)$, will also be called (self) *information* (content) $I(x) = b(x) = -L(x)$. (For an axiomatic approach, properties of information, and the definition of related quantities see, for example, [23, 29, 103].) Accordingly, Eq. (B.11) can be written

$$I(x) = -\ln p(x), \quad p(x) = e^{-I(x)}. \tag{B.18}$$

Like any random variable the bit number or (self) information $b(x)$ can be averaged over the whole set $\mathcal{X}$, over a subset $A \subseteq \mathcal{X}$, or conditioned on $Y = y$. One finds for the average or first moment of the bit number or (self) information b the *average (self) information* I_X

$$I_X(\mathcal{X}) = \langle b(x) \rangle_{\mathcal{X}} = -\langle \ln p(x) \rangle_{\mathcal{X}} \tag{B.19}$$

$$I_X(A) = \langle b(x) \rangle_A = -\langle \ln p(x) \rangle_A \tag{B.20}$$

$$I_X(y) = \langle b(x, y) \rangle_{\mathcal{X}|y} = -\langle \ln p(x, y) \rangle_{\mathcal{X}|y} \tag{B.21}$$

In Eq. (B.20) we used $b(x,A) = b(x)$ for $x \in A$ and $p(x|A)\ln p(x,A) = 0\ln 0 = 0$ for $x \notin A$ and in Eq. (B.21) we allowed y–dependent $p(x,y)$. It follows in accordance with the definition from Eq. (B.20) for $A = \{x\}$ that $I_X(A = \{x\}) = I_X(x) = I(x) = b(x)$ or analogously $I_X(Y = X = x) = b(x,x) = b(x)$ from Eq. (B.21). While I and b coincide on events $x \in \mathcal{X}$, in general the difference between bit number and average information, i.e., the difference between transformed probability and expectations of the corresponding canonical random variable, is given by

$$I_X(y) - b(y) = \langle b(x,y)\rangle_{x|y} - b(y) = -\langle \ln p(x,y)\rangle_{x|y} + \ln p(y)$$

$$= -\left\langle \ln \frac{p(x,y)}{p(y)} \right\rangle_{x|y} = -\langle \ln p(x|y)\rangle_{x|y} = \langle b(x|y)\rangle_{x|y}\,. \tag{B.22}$$

In terms of the *entropy*

$$H_X(y) = \langle b(x|y)\rangle_{x|y} \tag{B.23}$$

this can be written

$$I_X(y) - b(y) = H_X(y), \tag{B.24}$$

including

$$I_X(A) - b(A) = H_X(A). \tag{B.25}$$

The relations $b(\mathcal{X}) = 0$ and $I_X(x) = b(x)$ yield the special cases

$$H_X(\mathcal{X}) = I_X(\mathcal{X}), \quad H_X(x) = 0. \tag{B.26}$$

B.3.3 Energy

In the same way as $b(x)$, $F(x)$ may also be interpreted as a random variable on $\mathcal{X}$ called *energy* $E(x) = F(x)$ (or, for example, Euclidian *action* in field theory). Hence, energy is the canonical random variable of the transformation $-\frac{1}{\beta}\ln(Z(\mathcal{X})p(x))$, and one can write for Eqs. (B.13)

$$E(x) = -\frac{1}{\beta}\ln Z(x), \quad Z(x) = Z(x|\beta) = e^{-\beta E(x)}, \quad p(x) = e^{-\beta(E(x)-F(\mathcal{X}))}. \tag{B.27}$$

In general βF can be decomposed into $\beta F = \sum_i \alpha_i E_i$. (This is the case, for example, in the grand canonical ensemble of statistical physics where, compared to the canonical ensemble, a component corresponding to the particle number is added.)

The analogue of average informations I are then averages of the energy $E(x) = F(x)$ called *average energies* E_X

$$E_X(\mathcal{X}) \;=\; \langle F(x)\rangle_x = -\frac{1}{\beta}\langle \ln Z(x)\rangle_x \tag{B.28}$$

$$E_X(A) = \langle F(x)\rangle_A = -\frac{1}{\beta}\langle \ln Z(x)\rangle_A , \tag{B.29}$$

$$E_X(y) = \langle F(x,y)\rangle_{\mathcal{X}|y} = -\frac{1}{\beta}\langle \ln Z(x,y)\rangle_{\mathcal{X}|y} . \tag{B.30}$$

It follows in accordance with the definition from Eq. (B.29) for $A = \{x\}$ that $E_X(A = \{x\}) = E_X(x) = E(x) = F(x)$ or analogously $E_X(Y = X = x) = F(x,x) = F(x)$ from Eq. (B.30). We will skip the subscripts X in the following. While free energy and energy coincide on events $x \in \mathcal{X}$ in general their difference is given by

$$\beta E(y) - \beta F(y) = -\langle \ln Z(x,y)\rangle_{\mathcal{X}|y} + \ln Z(y)$$

$$= -\left\langle \ln \frac{Z(x,y)}{Z(y)} \right\rangle_{\mathcal{X}|y} = -\left\langle \ln \left(p(x|y) \frac{Z(\mathcal{Y})}{Z(\mathcal{X},\mathcal{Y})} \right) \right\rangle_{\mathcal{X}|y} . \tag{B.31}$$

Choosing

$$F(y) = -\frac{1}{\beta}\ln Z(y) = -\frac{1}{\beta}\ln \int dx\, e^{-\beta F(x,y)}, \tag{B.32}$$

i.e.,

$$Z(y) = Z(\mathcal{X},y) = \int dx\, Z(x,y) \tag{B.33}$$

it follows

$$Z(\mathcal{Y}) = \int dy\, Z(y) = \int dx \int dy\, Z(x,y) = Z(\mathcal{X},\mathcal{Y}). \tag{B.34}$$

Therefore

$$-\langle \ln Z(x,y)\rangle_{\mathcal{X}|y} + \ln Z(y) = -\langle \ln p(x|y)\rangle_{\mathcal{X}|y} \tag{B.35}$$

and

$$\langle F(x,y)\rangle_{\mathcal{X}|y} - F(y) = \frac{1}{\beta}\langle b(x|y)\rangle_{\mathcal{X}|y} . \tag{B.36}$$

Assuming $F(x|y)$ here to be defined as

$$F(x|y) = -\frac{1}{\beta}\ln \frac{Z(x,y)}{Z(y)} = -\frac{1}{\beta}\ln \frac{e^{-\beta F(x,y)}}{e^{-\beta F(y)}} = -\frac{1}{\beta}\ln \frac{e^{-\beta F(x,y)}}{\int dx\, e^{-\beta F(x,y)}}, \tag{B.37}$$

i.e.,

$$Z(x|y) = \frac{Z(x,y)}{Z(y)} = p(x|y), \tag{B.38}$$

one can also write for Eq. (B.36)

$$\langle F(x,y)\rangle_{\mathcal{X}|y} - F(y) = \langle F(x|y)\rangle_{\mathcal{X}|y} , \tag{B.39}$$

Transformed probability	Average of canonical random variable	Difference
bit number $b(y) = -\ln p(y)$	(self) information $I(y) = -\langle \ln p(x) \rangle_{x\|y}$	entropy $H(y) = -\langle \ln p(x\|y) \rangle_{x\|y}$ $= I(y) - b(y)$
free energy $F(y) = -\frac{1}{\beta} \ln Z(y)$	energy $E(y) = -\frac{1}{\beta} \langle \ln Z(x) \rangle_{x\|y}$	entropy $H(y) = -\langle \ln p(x\|y) \rangle_{x\|y}$ $= \beta(E(y) - F(y))$

Table B.1: Transformations of probability and averages of the corresponding canonical random variables.

parallelizing the corresponding equation(B.22) for bit numbers or (self) informations b. Using the definitions of entropy $H(y) = \langle b(x|y) \rangle_{x|y}$ and energy $E(y) = \langle F(x,y) \rangle_{x|y}$ Eq. (B.36) states that the β-scaled difference between energy and free energy is the entropy

$$\beta E(y) - \beta F(y) = H(y). \tag{B.40}$$

This is a generalization of the well known relation of statistical physics

$$\beta E(\mathcal{X}) - \beta F(\mathcal{X}) = H(\mathcal{X}), \tag{B.41}$$

where the argument $\mathcal{X}$ is usually skipped.

The Table B.1 summarizes some of the relations. (Recall that the variable y in the table can be replaced by $x \in \mathcal{X}$ or $A \subseteq \mathcal{X}$ and note that for $Z = \beta = 1$ free energy and energy become identical to bit number and information.)

B.4 Temperature and external fields

Now we look at the role of the parameter β. Its inverse $T = 1/\beta$ is called *temperature* in statistical mechanics. The parameter β can also be interpreted as an *external source or field* coupling to the conjugated random variable energy. The energy can thereby be subdivided into several components E_i with corresponding conjugated β_i. One β_i, for example, can be proportional to a magnetic field (or chemical potential, pressure, $\cdots$) coupling to a magnetic moment $E_i = M$ (or particle number, volume, $\cdots$). Also the calculation of moments or cumulants of other random variables is often facilitated by introducing an external source coupling to that variable.

We will discuss the following roles of β:

1. β is the Lagrange parameter determining the expectation of the energy $\langle E \rangle = \int dx\, p(x) E(x)$. Its variation defines an exponential family.

2. β is a homotopy parameter used by annealing methods, interpolating between easy and difficult to solve problems.

3. β represents an external source or field coupling to the energy. The cumulants of E can be obtained as responses to a changing external field, i.e., as derivatives of $\ln Z$ with respect to β. For example, $\langle E \rangle_{\mathcal{X}} = -(\partial/\partial\beta) \ln Z(\mathcal{X})$.

B.4.1 Maximum entropy and Boltzmann–Gibbs distributions

It is well known that minimizing the entropy and fixing normalization condition $\int dx\, p(x) = 1$ and expectations $\int dx\, p(x) E_i(x) = E_i(\mathcal{X})$ by the Lagrange multiplier method yield Boltzmann–Gibbs distributions (or generalized canonical ensembles). Indeed, adding the constraints with Lagrange multipliers α_i and setting to zero the functional derivative of

$$H(y) - \sum_{i=1} \alpha_i E_i(y) - \alpha_0 \langle 1 \rangle_{\mathcal{X}|y} \tag{B.42}$$

$$= - \int dx \left(p(x|y) \ln p(x|y) - \sum_{i=1} \alpha_i E_i(x,y) p(x|y) - \alpha_0 p(x|y) \right), \tag{B.43}$$

with respect to $p(x|y)$

$$0 = \frac{\delta}{\delta p(x|y)} \left(H(y) - \sum_{i=1} \alpha_i E_i(y) - \alpha_0 \langle 1 \rangle_{\mathcal{X}|y} \right) \tag{B.44}$$

$$= - \ln p(x|y) - 1 - \sum_{i=1} \alpha_i E_i(x,y) - \alpha_0, \tag{B.45}$$

one finds

$$p(x|y) = e^{-\sum_{i=1} \alpha_i E_i(x,y) - \alpha_0 - 1} = \frac{e^{-\beta F(x|y)}}{Z(\mathcal{X}|y)}, \tag{B.46}$$

with

$$\sum_{i=1} \alpha_i E_i(x,y) = \beta E(x,y) = \beta F(x|y) \quad \text{and} \quad Z(\mathcal{X}|y) = e^{\alpha_0 + 1}. \tag{B.47}$$

For $p(x|y) = p(x)$ this gives Eq. (B.27). Thus any probability distribution $p(x)$ can be seen as a result of a maximum entropy procedure with normalization constraint and fixed expectation of the energy $E(x)$.

B.4.2 Annealing methods

The Lagrange multiplier β, respectively the temperature $T = 1/\beta$, determines the average energy. Introduction of several Lagrange multipliers allows the fixation of several components E_i of $E(x)$. Varying β defines an exponential family with canonical parameter β and and canonical statistic E. In the high temperature limit $T \to \infty$, i.e., $\beta \to 0$, all $p(x)$ become equal. In the low temperature limit $T \to 0$, i.e., $\beta \to \infty$, only events x^* with maximal $p(x^*) \geq p(x)$, $\forall x \in \mathcal{X}$ survive, while all other events x with $p(x) < p(x^*)$, $\exists x^* \in \mathcal{X}$ are damped exponentially with decreasing temperature. The temperature dependency is used by annealing methods which are specific realizations of general homotopy or parameter continuation methods and very important in practice [261,338,428,533]. They solve a difficult (e.g., multimodal) problem at finite or zero temperature by beginning with an easier (e.g., convex) high temperature problem and then slowly decrease the temperature.

B.4.3 Generating functions

Moments or cumulants of random variables can often be conveniently calculated by the use of generating functions. The nth *moment* M_n of a random function $g(x)$, with $E(x)$ being a special case, is the expectation of its nth power

$$M_n(g) = \langle g^n(x) \rangle_x \,, \quad \text{e.g.,} \quad M_n(E) = \langle E^n(x) \rangle_x \,. \tag{B.48}$$

For a vector valued function g with components g_i (e.g., E_i), $i \in \mathcal{I}$ the moments become the functions (*unconnected correlation functions*)

$$M_{i_1,i_2,\cdots,i_n} = \langle g_{i_1}(x) g_{i_2}(x) \cdots g_{i_n}(x) \rangle_x \,. \tag{B.49}$$

The *cumulants* (or *connected correlation functions*) are given by [142,351]

$$C_{i_1,i_2,\cdots,i_n} = \sum_{\mathcal{P}} (-1)^{m-1} (m-1)!\, M_{j_1,j_2,\cdots,j_{p_1}} M_{k_1,k_2,\cdots,k_{p_2}} \cdots M_{l_1,l_2,\cdots,l_{p_m}} \tag{B.50}$$

with inverse

$$M_{i_1,i_2,\cdots,i_n} = \sum_{\mathcal{P}} C_{j_1,j_2,\cdots,j_{p_1}} C_{k_1,k_2,\cdots,k_{p_2}} \cdots C_{l_1,l_2,\cdots,l_{p_m}} \tag{B.51}$$

where $\mathcal{P}$ denotes a partition of the n indexes into non–empty subsets and m is the number of factors in the summand and one takes $C_0 = 0$. For a small number m of components i moments and cumulants may be more conveniently indexed by 'occupation numbers' n_i

$$M_{(n_1,n_2,\cdots,n_m)} = \langle g_1^{n_1}(x) g_2^{n_2}(x) \cdots g_n^{n_m}(x) \rangle_x \,. \tag{B.52}$$

For scalar function g, i.e., in the one component case $\mathcal{I} = \{1\}$, we will write the nth moment $M_{1,1,\cdots,1} = M_{(n)} = M_n$ and nth cumulant $C_{1,1,\cdots,1} = C_{(n)} = C_n$, skipping the bracket for the sake of simplicity. Hence, for a scalar g

$$M_0 = 1, \quad M_1 = C_1, \quad M_2 = C_2 + (C_1)^2, \quad M_3 = C_3 + 3C_2C_1 + (C_1)^3, \tag{B.53}$$

$$C_0 = 0, \quad C_1 = M_1, \quad C_2 = M_2 - (M_1)^2, \quad C_3 = M_3 - 3M_2M_1 + 2(M_1)^3, \tag{B.54}$$

where the second cumulant is the well known variance. Unlike moments, the cumulants are additive for independent subsystems, i.e., $p(x_1, x_2) = p(x_1)p(x_2) \Rightarrow C_n(p(x_1, x_2)) = C_n(p(x_1)) + C_n(p(x_2))$. Another significant property of cumulants is the possibility to set consistently $C_n = 0$ for all $n > 2$ (for Gaussian distributions), which is not possible for moments. If, however, one $C_n \neq 0$ for $n > 2$ then automatically an infinite number of other C_m do also not vanish (see, for example, [142]). For a multidimensional Gaussian distribution with vanishing means $M_i = 0$ Eq. (B.51) reduces to a sum over two–point connected correlation functions (or propagators) C_{ij}

$$M_{i_1,i_2,\cdots,i_{2n}} = \sum_{\text{Pairings}} C_{j_1,k_1} C_{j_2,k_2} \cdots C_{j_n,k_n}. \tag{B.55}$$

This relation is known as Wick's theorem.

For a scalar function $g(x)$ the *moment generating function* is given by

$$\begin{aligned} M(\gamma) &= \frac{Z(\gamma)}{Z} = \left\langle e^{\gamma g(x)} \right\rangle_x = \int dx\, p(x) e^{\gamma g(x)} \\ &= \frac{1}{Z} \int dx\, e^{-\beta E(x) + \gamma g(x)} = \sum_{n=0}^{\infty} \frac{\gamma^n}{n!} M_n, \end{aligned} \tag{B.56}$$

with

$$Z(\gamma) = \int dx\, e^{-\beta E(x) + \gamma g(x)}. \tag{B.57}$$

For vector g, with discrete or continuous index set $\mathcal{I}$, one has $\gamma g = \sum_i \gamma_i g_i$ (or $\int di$ for continuous i) and $\gamma^n M_n$ has to be understood as $\langle (\sum_i \gamma_i g_i(x))^n \rangle_x$, i.e.,

$$\begin{aligned} M(\gamma) &= \left\langle e^{\sum_i^m \gamma_i g_i(x)} \right\rangle_x = \sum_{n=0}^{\infty} \frac{1}{n!} \left\langle \left(\sum_i^m \gamma_i g_i(x) \right)^n \right\rangle_x \\ &= \sum_{n=0}^{\infty} \sum_{n_1,\ldots,n_m}^{n} \delta(\sum_i^m n_i - n) \frac{\gamma_1^{n_1} \cdots \gamma_n^{n_m}}{n_1! n_2! \cdots n_m!} M_{(n_1,n_2,\cdots,n_m)}, \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{i_1,\ldots,i_n}^{m} \gamma_{i_1} \cdots \gamma_{i_n} M_{i_1,i_2,\cdots,i_n}. \end{aligned} \tag{B.58}$$

It is easy to verify that the moments can be obtained by

$$\frac{\partial^n}{\partial\gamma^n}M(\gamma)\Big|_{\gamma=0} = \frac{\partial^n}{\partial\gamma^n}\left\langle e^{\gamma g(x)}\right\rangle_x\Big|_{\gamma=0}$$

$$= \left\langle \frac{\partial^n}{\partial\gamma^n}e^{\gamma g(x)}\right\rangle_x\Big|_{\gamma=0} = \left\langle g^n(x)e^{\gamma g(x)}\right\rangle_x\Big|_{\gamma=0} = \langle g^n(x)\rangle_x = M_n(g), \quad \text{(B.59)}$$

or in the multidimensional case

$$\frac{\partial^n}{\partial\gamma_{i_1}\cdots\partial\gamma_{i_n}}M(\gamma)\Big|_{\gamma=0} = \frac{\partial^n}{\partial\gamma_{i_1}\cdots\partial\gamma_{i_n}}\left\langle e^{\sum_i \gamma_i g_i(x)}\right\rangle_x\Big|_{\gamma=0} = M_{i_1,i_2,\cdots,i_n}(g). \quad \text{(B.60)}$$

The cumulants are generated by differentiating the logarithm $\ln M(\gamma) = C(\gamma)$

$$\frac{\partial^n}{\partial\gamma^n}C(\gamma)\Big|_{\gamma=0} = \frac{\partial^n}{\partial\gamma^n}\ln\left\langle e^{\gamma g(x)}\right\rangle_x\Big|_{\gamma=0} = C_n(g) \quad \text{(B.61)}$$

$$\frac{\partial^n}{\partial\gamma_{i_1}\cdots\partial\gamma_{i_n}}C(\gamma)\Big|_{\gamma=0} = \frac{\partial^n}{\partial\gamma_{i_1}\cdots\partial\gamma_{i_n}}\ln\left\langle e^{\sum_i \gamma_i g_i(x)}\right\rangle_x\Big|_{\gamma=0} = C_{i_1,i_2,\cdots,i_n}(g). \quad \text{(B.62)}$$

Hence, $C(\gamma)$ is the *cumulant generating function* with Taylor expansion around $\gamma = 0$

$$C(\gamma) = \ln M(\gamma) = \ln\left\langle e^{\gamma g(x)}\right\rangle_x = \sum_{n=0}^{\infty}\frac{\gamma^n}{n!}C_n(g), \quad \text{(B.63)}$$

or in the multidimensional case

$$C(\gamma) = \ln\left\langle e^{\sum_i \gamma_i g_i(x)}\right\rangle_x = \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{i_1,\ldots,i_n}\gamma_{i_1}\cdots\gamma_{i_n}C_{i_1,i_2,\cdots,i_n}. \quad \text{(B.64)}$$

Analogous to β the parameter γ can be thought of as an external source or field (e.g., a magnetic field) coupling to g. Even if a field γ is not present in 'reality' the cumulants can still be calculated as derivatives at $\gamma = 0$. If a nonzero field γ_0 is present we can incorporate $\gamma_0 g$ in a new energy $-\beta E = -\beta' E' + \gamma_0 g$, replacing a given $\beta' E'$ and proceeding as before. Because it is sometimes useful, especially in obtaining the Legendre transform of C, we will give in the following some of the formulae explicitly for nonzero field γ_0. Assuming that derivatives and integration can be interchanged one has

$$\frac{\partial^n}{\partial\gamma^n}M(\gamma)\Big|_{\gamma=\gamma_0} = \frac{\partial^n}{\partial\gamma^n}\left\langle e^{\gamma g(x)}\right\rangle_x\Big|_{\gamma=\gamma_0} = \left\langle g^n(x)e^{\gamma_0 g(x)}\right\rangle_x = \frac{Z(\gamma_0)}{Z}M_{n,\gamma_0}(g). \quad \text{(B.65)}$$

Including the γ_0–term in the expectation $\langle\cdots\rangle_{x|\gamma_0}$ we find for

$$M_{\gamma_0}(\gamma) = \frac{Z(\gamma)}{Z(\gamma_0)} = \frac{M(\gamma)}{M(\gamma_0)} = \left\langle e^{(\gamma-\gamma_0)g} \right\rangle_{x|\gamma_0} \tag{B.66}$$

as derivatives

$$\frac{\partial^n}{\partial\gamma^n} M_{\gamma_0}(\gamma)\Big|_{\gamma=\gamma_0} = \frac{\partial^n}{\partial\gamma^n} \frac{Z(\gamma)}{Z(\gamma_0)}\bigg|_{\gamma=\gamma_0} = \frac{\partial^n}{\partial\gamma^n} \left(\frac{\int dx\, e^{-\beta E(x)+\gamma g(x)}}{\int dx\, e^{-\beta E(x)+\gamma_0 g(x)}} \right) \bigg|_{\gamma=\gamma_0}$$

$$= \left\langle g^n e^{(\gamma-\gamma_0)g} \right\rangle_{x|\gamma_0} \Big|_{\gamma=\gamma_0} = \langle g^n(x) \rangle_{x|\gamma_0} = M_{n,\gamma_0}(g). \tag{B.67}$$

Cumulants for nonzero fields are generated by $C_{\gamma_0} = \ln M_{\gamma_0}$

$$\frac{\partial^n}{\partial\gamma^n} C_{\gamma_0}(\gamma)\Big|_{\gamma=\gamma_1} = \frac{\partial^n}{\partial\gamma^n} \ln \left\langle e^{(\gamma-\gamma_0)g(x)} \right\rangle_{x|\gamma_0} \Bigg|_{\gamma=\gamma_1} = C_{n,\gamma_1}(g), \tag{B.68}$$

where $\gamma_1 = \gamma_0$ is possible. Because an additive constant does not change the derivatives of a cumulant generating function $C(\gamma) = \ln M(\gamma) = \ln M_{\gamma_0}(\gamma) + \ln Z(\gamma_0) - \ln Z(\gamma) = \ln M_{\gamma_0}(\gamma) + \ln M(\gamma_0)$ can also be used as cumulant generating function for $\gamma_0 \neq 0$. The expansion of the generating functions M_{γ_0} and C_{γ_0} around γ_0 in powers of $(\gamma - \gamma_0)$ becomes

$$M_{\gamma_0}(\gamma) = \left\langle e^{(\gamma-\gamma_0)g(x)} \right\rangle_{x|\gamma_0} = \sum_{n=0}^{\infty} \frac{(\gamma-\gamma_0)^n}{n!} M_{n,\gamma_0}(g), \tag{B.69}$$

$$C_{\gamma_0}(\gamma) = \ln M(\gamma) = \ln \left\langle e^{(\gamma-\gamma_0)g(x)} \right\rangle_{x|\gamma_0}$$

$$= \sum_{n=0}^{\infty} \frac{(\gamma-\gamma_0)^n}{n!} C_{n,\gamma_0}(g) = \sum_{n=0}^{\infty} \frac{(\gamma-\gamma_1)^n}{n!} C_{n,\gamma_1}(g), \tag{B.70}$$

and analogously for the multidimensional case. Moments and cumulants for different fields γ_0 and γ_1 are related according to

$$\begin{aligned} M_{n,\gamma_0} &= \frac{Z(\gamma_1)}{Z(\gamma_0)} \frac{\partial^n}{\partial\gamma^n} M_{\gamma_1}(\gamma)\Big|_{\gamma=\gamma_0} \\ &= \frac{\left\langle g^n e^{(\gamma_0-\gamma_1)g} \right\rangle_{x|\gamma_1}}{\left\langle e^{(\gamma_0-\gamma_1)g} \right\rangle_{x|\gamma_1}} \\ &= \frac{\sum_{m=0}^{\infty} \frac{(\gamma_0-\gamma_1)^m}{m!} M_{m+n,\gamma_1}}{\sum_{k=0}^{\infty} \frac{(\gamma_0-\gamma_1)^k}{k!} M_{k,\gamma_1}}, \end{aligned} \tag{B.71}$$

and

$$C_{n,\gamma_0} = \frac{\partial^n}{\partial \gamma^n} C_{\gamma_1}(\gamma)\Big|_{\gamma=\gamma_0} = \sum_{m=0}^{\infty} \frac{(\gamma_0 - \gamma_1)^m}{m!} C_{m+n,\gamma_1}, \tag{B.72}$$

i.e., for the difference

$$\Delta C_n(\gamma_0, \gamma_1) = C_{n,\gamma_0} - C_{n,\gamma_1} = \sum_{m=1}^{\infty} \frac{(\gamma_0 - \gamma_1)^m}{m!} C_{m+n,\gamma_1}. \tag{B.73}$$

Because inverse temperature β can be regarded as a special nonzero field, the moments and cumulants of the energy can be obtained by

$$\frac{\partial^n}{\partial(-\beta)^n} Z = Z M_n(E) \tag{B.74}$$

and

$$\frac{\partial^n}{\partial(-\beta)^n} \ln Z = C_n(E), \tag{B.75}$$

like

$$\frac{\partial^2}{\partial(-\beta)^2} \ln Z = C_2(E) = \left\langle E^2(x) \right\rangle_x - \left\langle E(x) \right\rangle_x^2 . \tag{B.76}$$

The generalization to the multidimensional case is analogous to Eqs. (B.58) and (B.60). Equations for the second cumulants connecting the derivative with respect to an external field (response, dissipation) with a variance (fluctuation) are also known under the name dissipation–fluctuation theorems.

Moments of a function g of x can be expressed by moments of $\Delta x = x - x_0$ by expanding $g(x)$ around x_0

$$g(x) = \sum_{n=0}^{\infty} \frac{(x - x_0)^n}{n!} \frac{\partial^n}{\partial x^n} g(x)\Big|_{x=x_0} = e^{\langle \Delta x | \nabla' \rangle} g(x')\Big|_{x'=x_0} = e^{\langle \Delta x | \nabla \rangle} g(x_0), \tag{B.77}$$

with gradient $\nabla' g(x') = (\partial/\partial x') g(x') = g^{(1)}(x')$ and analogously $(\nabla')^n g(x') = (\partial^n/\partial(x')^n) g(x') = g^{(n)}(x')$. We understand here and in the following that the expression $e^{\langle \Delta x | \nabla \rangle}$, or $\langle \Delta x | \nabla \rangle^n$, is to be 'normal ordered', meaning that the derivatives act only to the right and not on Δx. Analogously in the multidimensional case, for example, $\nabla^2 = \Delta$ creates the Hessian matrix. This yields,

$$M_n = \left\langle g(x) \right\rangle_x = \left\langle g(x_0) + \Delta x\, g^{(1)}(x_0) + \frac{(\Delta x)^2}{2} g^{(2)}(x_0) + \cdots \right\rangle_x \tag{B.78}$$

$$= \left\langle e^{\langle \Delta x | \nabla \rangle} g(x_0) \right\rangle_x = g(x_0) + \left\langle \Delta x \right\rangle_x g^{(1)}(x_0) + \frac{\left\langle (\Delta x)^2 \right\rangle_x}{2} g^{(2)}(x_0) + \cdots . \tag{B.79}$$

We have seen that an expansion in moments or cumulants depends on the the choice of γ_0 or, equivalently, on the splitting of the x–dependent terms in one term $-\beta E(x)$ which defines an expectation $\langle\cdots\rangle_x$ and a field term $\gamma g(x)$. Thus, there is a practically important freedom in choosing different moment or cumulant expansions as approximations for the same exponential. Assume, for example, that moments for $p'(x) \propto e^{-\beta' E'(x)}$ can be easily calculated. For $-\beta E = -\beta' E' + \gamma g(x)$ we can then write Eq. (B.71) for an expectation $\langle\cdots\rangle_x$ of a function $q(x)$ under $p(x) \propto e^{-\beta E(x)}$

$$\langle q(x)\rangle_x = \langle q e^{\gamma g}\rangle'_x \frac{Z'}{Z} = \frac{\langle q e^{\gamma g}\rangle'_x}{\langle e^{\gamma g}\rangle'_x} = \frac{\langle q\,(1+\gamma g+\cdots)\rangle'_x}{\langle 1+\gamma g+\cdots\rangle'_x}, \tag{B.80}$$

with $Z' = \int dx\, e^{-\beta' E'(x)}$ and $\langle\cdots\rangle'_x = \int dx\, p'(x)\cdots$. By expansion around x_0 this can also be expressed in terms of moments of Δx

$$\langle q(x)\rangle_x = \frac{e^{\gamma g(x_0)}\left(q(x_0) + \langle\Delta x\rangle'_{x,(\gamma)}\left(q^{(1)}(x_0) + q(x_0)\gamma g^{(1)}(x_0)\right) + \cdots\right)}{e^{\gamma g(x_0)}\left(1 + \langle\Delta x\rangle'_{x,(\gamma)}\,\gamma g^{(1)}(x_0) + \cdots\right)}, \tag{B.81}$$

where the prefactor cancels. This is the basis of saddle point approximation (or loop expansion) (see, for example, Section 2.1.6) and also the basis of importance sampling in Monte Carlo calculations [47,49,150,363].

To obtain equations which go beyond a cumulant expansion it is useful to consider the Legendre transform of $C(\gamma)$. The *Legendre transform* (or *effective action*) $\Gamma(\phi)$ of $C(\gamma)$ is defined by requiring [165,425,541]

$$\Gamma(\phi) + C(\gamma) - \sum_i \gamma_i \phi_i, \tag{B.82}$$

to be stationary with respect to variations of fields γ_i (coupling to g_i) at ϕ fixed. (We may remark here that $C(\gamma)$, and thus its Legendre transform Γ, depends on the choice of g_i. A typical case is $g_i = x_i$.) This means

$$\frac{\partial C(\gamma)}{\partial \gamma_i} = \phi_i = \langle g\rangle_{x|\gamma} = C_{i,\gamma}, \tag{B.83}$$

using also that C is the cumulant generating function. Using the chain rule one finds by differentiating functional (B.82) with respect to ϕ_i

$$\frac{\partial \Gamma(\phi)}{\partial \phi_i} = -\sum_i \frac{\partial C(\gamma)}{\partial \gamma_i}\frac{\partial \gamma_i}{\partial \phi_i} + \gamma_i + \sum_i \phi_i \frac{\partial \gamma_i}{\partial \phi_i}, \tag{B.84}$$

and thus with Eq. (B.83)

$$\frac{\partial \Gamma(\phi)}{\partial \phi_i} = \gamma_i. \tag{B.85}$$

Now set

$$\Gamma(\phi) = \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{i_1,\cdots,i_n} \Gamma_{i_1,\cdots,i_n} \Delta C_{i_1} \cdots \Delta C_{i_n}, \tag{B.86}$$

with $\Delta C_i = C_{i,\gamma} - C_i = \phi_i - C_i$. This defines the *(proper) vertex functions* $\Gamma_{i_1,\cdots,i_n}$ for which

$$\Gamma_{i_1,\cdots,i_n} = \left.\frac{\partial^n \Gamma}{\partial\phi_{i_1}\cdots\partial\phi_{i_n}}\right|_{\gamma=0} = \left.\frac{\partial^{n-1}\gamma_{i_n}(\phi)}{\partial\phi_{i_1}\cdots\partial\phi_{i_{n-1}}}\right|_{\gamma=0}, \tag{B.87}$$

using Eq. (B.85) and noting that $\Delta C = 0$ if $\gamma = 0$. Inverting the multidimensional version of Eq. (B.73) (setting $\gamma_1 = 0$) to obtain $\gamma_i(\phi)$ in terms of ΔC_i, and defining *amputated correlation functions*

$$A_{i_1,\cdots,i_n} = \sum_{j_1,\cdots j_n} C^{-1}_{i_1,j_1} \cdots C^{-1}_{n_1,j_n} C_{j_1,\cdots,j_n}, \tag{B.88}$$

one finds

$$\Gamma_{i_1} = 0, \tag{B.89}$$

$$\Gamma_{i_1,i_2} = C^{-1}_{i_1,i_2}, \tag{B.90}$$

$$\Gamma_{i_1,i_2,i_3} = -A_{i_1,i_2,i_3}, \tag{B.91}$$

$$\begin{aligned}\Gamma_{i_1,i_2,i_3,i_4} &= -A_{i_1,i_2,i_3,i_4} + \sum_{i,j} (A_{i_1,i_2,i} C_{i,j} A_{j,i_3,i_4} \\ &\quad + A_{i_1,i_3,i} C_{i,j} A_{j,i_2,i_4} + A_{i_1,i_4,i} C_{i,j} A_{j,i_2,i_3}),\end{aligned} \tag{B.92}$$

$\cdots$

These equations can easily be inverted

$$C_{i_1,i_2} = \Gamma^{-1}_{i_1,i_2}, \tag{B.93}$$

$$A_{i_1,i_2,i_3} = -\Gamma_{i_1,i_2,i_3}, \tag{B.94}$$

$$\begin{aligned}A_{i_1,i_2,i_3,i_4} &= -\Gamma_{i_1,i_2,i_3,i_4} + \sum_{i,j} (\Gamma_{i_1,i_2,i} C_{i,j} \Gamma_{j,i_3,i_4} \\ &\quad + \Gamma_{i_1,i_3,i} C_{i,j} \Gamma_{j,i_2,i_4} + \Gamma_{i_1,i_4,i} C_{i,j} \Gamma_{j,i_2,i_3}),\end{aligned} \tag{B.95}$$

$\cdots$

Often it is only possible to find (full or perturbed) vertex functions $\Gamma_{i_1,\cdots,i_n}$ by expanding around known vertex functions $\Gamma^0_{i_1,\cdots,i_n}$ for a solvable (reference or unperturbed) system. For example, Eq. (B.90) can be reformulated in terms of the 'self energy' being the difference between perturbed and unperturbed two point vertex functions

$$(\Sigma)_{i_1,i_2} = \Gamma_{i_1,i_2} - \Gamma^0_{i_1,i_2}. \tag{B.96}$$

generating function	derivatives
$Z(\gamma) = \int dx\, e^{-\beta E(x)+\gamma g(x)}$	$\frac{\partial^n}{\partial\gamma^n} Z(\gamma)\Big\vert_{\gamma=0} = Z M_n(g)$
$\ln Z(\gamma) = \ln \int dx\, e^{-\beta E(x)+\gamma g(x)}$	$\frac{\partial^n}{\partial\gamma^n} \ln Z(\gamma)\Big\vert_{\gamma=0} = C_n(g)$
$M(\gamma) = \frac{Z(\gamma)}{Z} = \langle e^{\gamma g(x)} \rangle_x$	$\frac{\partial^n}{\partial\gamma^n} M(\gamma)\Big\vert_{\gamma=0} = M_n(g)$
$C(\gamma) = \ln M(\gamma) = \ln \langle e^{\gamma g(x)} \rangle_x$	$\frac{\partial^n}{\partial\gamma^n} C(\gamma)\Big\vert_{\gamma=0} = C_n(g)$

Table B.2: Generating functions for zero field.

This results in

$$C_{i_1,i_2} = C^0_{i_1,i_2} - \sum_{j,k} C^0_{i_1,j} \, (\Sigma)_{j,k} \, C_{k,i_2}, \tag{B.97}$$

which expresses the full connected two point correlation function C_{i_1,i_2} in terms of an unperturbed $C^0_{i_1,i_2}$. An approximation $(\hat{\Sigma})$ for the self energy can be obtained, for example, by perturbation theory with respect to the unperturbed reference system. Then a corresponding self-consistent solution $\hat{C}_{i_1,i_2}$ can be found by iteration according to

$$\hat{C}_{i_1,i_2} = C^0_{i_1,i_2} - \sum_{j,k} C^0_{i_1,j} (\hat{\Sigma})_{j,k} C^0_{k,i_2} + \sum_{j,k,l,m} C^0_{i_1,j} (\hat{\Sigma})_{j,k} C^0_{k,l} (\hat{\Sigma})_{l,m} C^0_{m,i_2}. \tag{B.98}$$

Table B.2 shows generating functions for zero field; analogously Table B.3 shows generating functions for nonzero field β or γ_0.

B.5 Conditional probabilities and disordered systems

We already discussed that it is often useful to look for conditions under which the energy is easy to specify and to combine the different possible conditions

generating function	derivatives
$Z = \int dx\, e^{-\beta E(x)}$	$\frac{\partial^n}{\partial(-\beta')^n} Z(\beta')\Big\|_{\beta'=\beta} = Z M_n(E)$
$\ln Z = \ln \int dx\, e^{-\beta E(x)}$	$\frac{\partial^n}{\partial(-\beta')^n} \ln Z(\beta')\Big\|_{\beta'=\beta} = C_n(E)$
$Z(\gamma) = \int dx\, e^{-\beta E(x)+\gamma g(x)}$	$\frac{\partial^n}{\partial\gamma^n} Z(\gamma)\Big\|_{\gamma=\gamma_0} = Z(\gamma_0) M_{n,\gamma_0}(g)$
$\ln Z(\gamma) = \ln \int dx\, e^{-\beta E(x)+\gamma g(x)}$	$\frac{\partial^n}{\partial\gamma^n} \ln Z(\gamma)\Big\|_{\gamma=\gamma_0} = C_{n,\gamma_0}(g)$
$M(\gamma) = \frac{Z(\gamma)}{Z} = \left\langle e^{\gamma g(x)} \right\rangle_x$	$\frac{\partial^n}{\partial\gamma^n} M(\gamma)\Big\|_{\gamma=\gamma_0} = M(\gamma_0) M_{n,\gamma_0}(g)$
$C(\gamma) = \ln M(\gamma) = \ln \left\langle e^{\gamma g(x)} \right\rangle_x$	$\frac{\partial^n}{\partial\gamma^n} C(\gamma)\Big\|_{\gamma=\gamma_0} = C_{n,\gamma_0}(g)$
$M_{\gamma_0}(\gamma) = \frac{Z(\gamma)}{Z(\gamma_0)} = \left\langle e^{(\gamma-\gamma_0)g(x)} \right\rangle_{x\|\gamma_0}$	$\frac{\partial^n}{\partial\gamma^n} M_{\gamma_0}(\gamma)\Big\|_{\gamma=\gamma_0} = M_{n,\gamma_0}(g)$
$C_{\gamma_0}(\gamma) = \ln M_{\gamma_0}(\gamma)$ $= \ln \left\langle e^{(\gamma-\gamma_0)g(x)} \right\rangle_{x\|\gamma_0}$	$\frac{\partial^n}{\partial\gamma^n} C_{\gamma_0}(\gamma)\Big\|_{\gamma=\gamma_1} = C_{n,\gamma_1}(g)$

Table B.3: Generating functions for nonzero fields.

in a second step to obtain the complete probability. Thus, a joint probability $p(x,y)$ can be specified either by a 'joint' (or annealed) energy function $E(x,y)$ with conjugated joint (or annealed) temperature $1/\beta$ or a conditional (or quenched) energy $E(x|y)$ with conjugated conditional (or quenched) temperatures $1/B(y)$ and a mixture energy $E(y)$ with mixture temperature $1/b$. Hence,

$$p(x,y) = \frac{e^{-\beta E(x,y)}}{Z(\mathcal{X},\mathcal{Y})} = p(y)p(x|y) = \frac{e^{-bE(y)}e^{-B(y)E(x|y)}}{Z(\mathcal{Y})Z(\mathcal{X}|y)}, \tag{B.99}$$

with

$$p(y) = e^{-bE(y)}/Z(\mathcal{Y}), \tag{B.100}$$

$$p(x|y) = e^{-B(y)E(x|y)}/Z(\mathcal{X}|y), \tag{B.101}$$

and

$$Z(\mathcal{X},\mathcal{Y}) = \int dx \int dy\, e^{-\beta E(x,y)}, \tag{B.102}$$

$$Z(\mathcal{Y}) = \int dy\, e^{-bE(y)}, \tag{B.103}$$

$$Z(\mathcal{X}|y) = \int dx\, e^{-B(y)E(x|y)}. \tag{B.104}$$

One may remark, that choosing $-\beta E(x,y) = -bE(y) - B(y)E(x|y)$ produces a joint probability

$$p'(x,y) = \frac{e^{-\beta E(x,y)}}{Z(\mathcal{X},\mathcal{Y})} = \frac{e^{-bE(y)-B(y)E(x|y)}}{\int dy\, e^{-bE(y)}Z(\mathcal{X}|y)}, \tag{B.105}$$

which is different from $p(x,y)$ of Eq. (B.99).

If interested only in variable x one integrates (marginalizes) over y. Working with joint energies this gives

$$p(x) = \int dy\, p(x,y) = \frac{\int dy\, e^{-\beta E(x,y)}}{Z(\mathcal{X},\mathcal{Y})} = \frac{e^{-\beta E(x)}}{Z(\mathcal{X})}, \tag{B.106}$$

whereas working with conditional energies yields

$$p(x) = \int dy\, p(x,y) = \int dy\, \frac{e^{-\ bE(y)}e^{-B(y)E(x|y)}}{Z(\mathcal{Y})Z(\mathcal{X}|y)} = \int dy\, p(y)\frac{e^{-B(y)E(x|y)}}{Z(\mathcal{X}|y)}. \tag{B.107}$$

In the formulation with joint probabilities $E(x,y)$ is the canonical variable. Its averages can be obtained by differentiation with respect to β

$$E_{ann} = E(\mathcal{X}) = \langle E(x)\rangle_{x} = \langle E(x,y)\rangle_{x,\mathcal{Y}} = -\frac{\partial}{\partial\beta}\ln Z(\mathcal{X}). \tag{B.108}$$

In the formulation with conditional probabilities the expectation of $E(x|y)$ can be obtained by differentiation with respect to $B(y)$

$$E_{quen} = E(\mathcal{X}|\mathcal{Y}) = \langle E(\mathcal{X}|y)\rangle_{\mathcal{Y}} = \langle E(x|y)\rangle_{x,\mathcal{Y}} = -\left\langle \frac{\partial}{\partial B(y)} \ln Z(\mathcal{X}|y)\right\rangle_{\mathcal{Y}} . \tag{B.109}$$

Even for y–independent $B(y) = B$ the averaging of $\ln Z(\mathcal{X}|y)$ remains

$$\langle E(x|y)\rangle_{x,\mathcal{Y}} = -\frac{\partial}{\partial B} \langle \ln Z(\mathcal{X}|y)\rangle_{\mathcal{Y}} . \tag{B.110}$$

For y–independent normalization $Z(\mathcal{X}|y) = Z(\mathcal{X})$ both approaches are equivalent and $\langle \ln Z(\mathcal{X}|y)\rangle_{\mathcal{Y}} = \ln Z(\mathcal{X})$. In general, however, the expectations E_{ann} and E_{quen} are different. Also, despite of the equality (*B*.99), the exponential families defined by varying the parameters β or B (or $B(y)$) are not the same. The conditional temperatures or fields $B(y)$ do not influence the distribution $p(y)$, while the joint temperature $1/\beta$ does.

In practice, for example, it may take some time after changing temperature or an external field until a stationary distribution $p(x, y)$ is reached. Assume the dynamic of y being much slower than that of x (e.g., lower energy barriers for x and higher energy barriers for y). In a magnetic substance, x may stand for fast adapting local spins and y for very slowly moving impurities. Then changing the physical temperature will on short time scales (or low temperatures) only change the distribution of the fast adapting spins x, while the slow impurities y remain quenched. Then the relevant physical temperature is the conditioned or quenched field $B(x)$. On a much longer time scale (or at high enough temperatures, i.e., in an 'annealed system') the impurities will also approach an equilibrium distribution. Then the physical temperature is a joint or annealed field β. In addition, ensemble averages for variables which reach a stationary distribution on short enough time scales are often measured as time averages under a (quasi–)ergodic dynamic. In contrast, averages over slow variables can in practice not be obtained as time averages and must be realized as ensemble averages.

Eq. (B.110) requires the calculation of a partition sum $Z(\mathcal{Y}, x)$ for every x. This is possible for a small number of different x values or if the x–dependence can be calculated analytically and the average over $\mathcal{X}$ be performed. In general, however, Eq. (B.110) is rather difficult to solve. One possibility to proceed is using the identity

$$\ln Z = \lim_{n\to 0} \frac{Z^n - 1}{n}, \tag{B.111}$$

which is verified by expanding

$$Z^n = e^{n \ln Z} = 1 + n \ln Z + \sum_{k=2} \frac{n^k}{k!} (\ln Z)^k. \tag{B.112}$$

Typically, the average over Z^n for integer n is easier to perform than over $\ln Z$. Because Z^n describes a system with n independent 'replicas' of the same system this approach is known under the name replica method [128,209,338,380]. Performing the average over x, however, results usually in a coupling between the different replicas. Also one has to be careful, because calculating Z^n for integer n does not uniquely determine the analytic continuation to $n \to 0$. For non–interacting disordered systems a supersymmetric approach (expressing the two–point correlation function as a product of two Gaussian integrals, one over commuting variables and another over anticommuting (Grassmann) variables) can avoid such difficulties of the replica approach [117,357].

This page intentionally left blank

Appendix C

Iteration procedures: Learning

C.1 Numerical solution of stationarity equations

This appendix discusses some numerical methods for solving the stationarity equations of nonparametric density estimation. While in principle any optimization algorithm which works well in the high dimensional spaces of interest can be used to find a maximum of the Bayesian posterior density, we will concentrate in the following on simple, easy to implement methods well suited for density estimation problems.

Due to the presence of the logarithmic data term $-\langle \ln P | N \rangle$ and the normalization constraint in Bayesian density estimation problems, the stationarity equations are in general nonlinear, even for Gaussian specific priors. An exception are Gaussian regression problems discussed in Section 3.8 for which the data term $-\langle \ln P | N \rangle$ becomes quadratic and the normalization constraint can be skipped. However, the nonlinearities arising from the data term are restricted to the finite number of training data points, and the nonlinear normalization constraint is, for Gaussian specific priors, numerically not very harmful. Clearly, more severe nonlinearities appear for general non–Gaussian specific priors, like in the presence of auxiliary fields (Section 5.7), for the non–Gaussian priors with non–quadratic potentials of Section 5.8, or for general nonlinear parameterizations $P(\xi)$. Those appear, for example, in the inverse problems with quantum mechanical likelihood in Chapter 7.

For nonlinear equations the stationarity conditions generally have to be solved by iteration. In the context of empirical learning, iteration procedures to minimize an error functional represent possible *learning algorithms.*

In Chapter 3 we have encountered stationarity equations

$$0 = \frac{\delta(-E(\phi))}{\delta\phi} = G(\phi) \tag{C.1}$$

for error functionals $E(\phi)$, with, for example, $\phi = L$ or $\phi = P$, written in the form

$$\mathbf{K}\phi = T. \tag{C.2}$$

with ϕ–dependent $T = G(\phi) + \mathbf{K}\phi$. For the stationarity Eqs. (3.28), (3.55), and (3.105) the operator $\mathbf{K}$ is a ϕ–independent inverse covariance of a Gaussian specific prior. For existing (and not too ill–conditioned) covariance $\mathbf{K}^{-1}$ of the prior process, a simple iteration scheme for Eq. (C.2) would be

$$\phi^{(i+1)} = \mathbf{K}^{-1}T(\phi^{(i)}), \tag{C.3}$$

starting from some initial guess $\phi^{(0)}$.

Note that writing the stationarity equation in the form (C.2) is no restriction of generality, if we allow a general ϕ–dependent T. It can also be convenient to work with a ϕ–dependent operator $\mathbf{K}$ in (C.2), like in Eq. (4.9) for general nonlinear parameterizations $P(\xi)$ or for the mixture models in Section 6. Clearly one always can choose an arbitrary invertible (and not too ill–conditioned) operator $\mathbf{A}$, define $T_{\mathbf{A}} = G(\phi) + \mathbf{A}\phi = T + (\mathbf{A} - \mathbf{K})\phi$, so that a stationarity equation $G(\phi) = T - \mathbf{K}\phi = T_{\mathbf{A}} - \mathbf{A}\phi = 0$ can be written $\mathbf{A}\phi = T_{\mathbf{A}}$. That equation can then be discretized and iterated with the (discretized) $\mathbf{A}^{-1}$ analogous to Eq. (C.3). To obtain a numerical iteration scheme we will choose a linear, positive definite *learning matrix* $\mathbf{A}$. The learning matrix may depend on ϕ and may also change during iteration.

To connect a stationarity equation given in form (C.2) to an arbitrary iteration scheme with a learning matrix $\mathbf{A}$ we define

$$\mathbf{B} = \mathbf{K} - \mathbf{A}, \quad \mathbf{B}_\eta = \mathbf{K} - \frac{1}{\eta}\mathbf{A}, \tag{C.4}$$

and split $\mathbf{K}$ into two parts

$$\mathbf{K} = \mathbf{A} + \mathbf{B} = \frac{1}{\eta}\mathbf{A} + \mathbf{B}_\eta, \tag{C.5}$$

where we introduced η for later convenience. Then we obtain from the stationarity equation (C.2)

$$\phi = \eta\mathbf{A}^{-1}(T - \mathbf{B}_\eta\phi). \tag{C.6}$$

To iterate we start by inserting an approximate solution $\phi^{(i)}$ on the right–hand side and obtain a new $\phi^{(i+1)}$ by calculating the left hand side. This can

be written in one of the following equivalent forms

$$\phi^{(i+1)} = \eta \mathbf{A}^{-1} \left(T^{(i)} - \mathbf{B}_\eta \phi^{(i)} \right) \tag{C.7}$$

$$= (1-\eta)\phi^{(i)} + \eta \mathbf{A}^{-1} \left(T^{(i)} - \mathbf{B} \phi^{(i)} \right) \tag{C.8}$$

$$= \phi^{(i)} + \eta \mathbf{A}^{-1} \left(T^{(i)} - \mathbf{K} \phi^{(i)} \right), \tag{C.9}$$

where η plays the role of a learning rate or step width, and $\mathbf{A}^{-1} = \left(\mathbf{A}^{(i)}\right)^{-1}$ may be iteration dependent. The update equations (C.7) – (C.9) can be written

$$\Delta\phi^{(i)} = \eta \mathbf{A}^{-1} G(\phi^{(i)}), \tag{C.10}$$

with $\Delta\phi^{(i)} = \phi^{(i+1)} - \phi^{(i)}$. Eq. (C.9) does not require the calculation of $\mathbf{B}$ or $\mathbf{B}_\eta$ so that instead of $\mathbf{A}$, directly $\mathbf{A}^{-1}$ can be given without the need to calculate its inverse. For example, operators approximating $\mathbf{K}^{-1}$ and being easy to calculate may be good choices for $\mathbf{A}^{-1}$.

For positive definite $\mathbf{A}$ (and thus also positive definite inverse $\mathbf{A}^{-1}$) convergence can be guaranteed, at least theoretically. Indeed, multiplying with $(1/\eta)\mathbf{A}$ and projecting on an infinitesimal $d\phi$ gives

$$\frac{1}{\eta} \left\langle d\phi^{(i)} \middle| \mathbf{A} \middle| \Delta\phi^{(i)} \right\rangle = \left\langle d\phi \middle| \frac{\delta(-E(\phi))}{\delta\phi}\Big|_{\phi=\phi^{(i)}} \right\rangle = d(-E(\phi^{(i)})). \tag{C.11}$$

In an infinitesimal neighborhood of $\phi^{(i)}$, where $\Delta\phi^{(i)}$ becomes equal to $d\phi^{(i)}$ in first order, the left–hand side is for positive (semi–)definite $\mathbf{A}$ larger than (or equal to) zero. This shows that at least for η small enough the log–posterior $-E(\phi)$ increases, i.e., the differential $dE(\phi)$ is smaller or equal to zero and the value of the error functional $E(\phi)$ decreases.

Stationarity equation (3.17) for minimizing $E(L)$ yields for (C.7, C.8, C.9),

$$L^{(i+1)} = \eta \mathbf{A}^{-1} \left(N - \mathbf{\Lambda}_X^{(i)} e^{L^{(i)}} - \mathbf{K} L^{(i)} + \frac{1}{\eta} \mathbf{A} L^{(i)} \right) \tag{C.12}$$

$$= (1-\eta) L^{(i)} + \eta \mathbf{A}^{-1} \left(N - \mathbf{\Lambda}_X^{(i)} e^{L^{(i)}} - \mathbf{K} L^{(i)} + \mathbf{A} L^{(i)} \right) \tag{C.13}$$

$$= L^{(i)} + \eta \mathbf{A}^{-1} \left(N - \mathbf{\Lambda}_X^{(i)} e^{L^{(i)}} - \mathbf{K} L^{(i)} \right). \tag{C.14}$$

The function $\Lambda_X^{(i)}$ is also unknown and is part of the variables we want to solve for. The normalization conditions provide the necessary additional equations, and the matrix $\mathbf{A}^{-1}$ can be extended to include the iteration procedure for Λ_X. For example, we can insert the stationarity equation for Λ_X in (C.14) to get

$$L^{(i+1)} = L^{(i)} + \eta \mathbf{A}^{-1} \left[N - e^{\mathbf{L}^{(i)}} (N_X - \mathbf{I}_X \mathbf{K} L^{(i)}) - \mathbf{K} L^{(i)} \right]. \tag{C.15}$$

If normalizing $L^{(i)}$ at each iteration this corresponds to an iteration procedure for $g = L + \ln Z_X$.

Similarly, for the functional $E(P)$ we have to solve (3.49) and obtain for (C.9),

$$\begin{aligned} P^{(i+1)} &= P^{(i)} + \eta \mathbf{A}^{-1} \left(T_P^{(i)} - \mathbf{K} P^{(i)} \right) \\ &= P^{(i)} + \eta \mathbf{A}^{-1} \left(\mathbf{P^{(i)}}^{-1} N - \Lambda_X^{(i)} - \mathbf{K} P^{(i)} \right) \\ &= P^{(i)} + \eta \mathbf{A}^{-1} \left(\mathbf{P^{(i)}}^{-1} N - N_X - \mathbf{I}_X \mathbf{P^{(i)}} \mathbf{K} P^{(i)} - \mathbf{K} P^{(i)} \right). \end{aligned} \tag{C.16}$$

When normalizing P at each iteration, this is equivalent to solving for $z = Z_X P$, and the update procedure for Λ_X can be varied.

C.2 Learning matrices

C.2.1 Learning algorithms for density estimation

There exists a variety of well developed numerical methods for unconstraint as well as for constraint optimization [28,40,134,168,190,191,398,402,410]. Popular examples include conjugate gradient, Newton, and quasi–Newton methods, like the variable metric methods DFP (Davidon–Fletcher–Powell) or BFGS (Broyden–Fletcher–Goldfarb–Shanno).

All of them correspond to the choice of specific, often iteration dependent, learning matrices $\mathbf{A}$ defining the learning algorithm. Possible simple choices are:

$$\mathbf{A} = \mathbf{I} \quad : \quad \text{Gradient} \tag{C.17}$$

$$\mathbf{A} = \mathbf{D} \quad : \quad \text{Jacobi} \tag{C.18}$$

$$\mathbf{A} = \mathbf{T} + \mathbf{D} \quad : \quad \text{Gauss–Seidel} \tag{C.19}$$

$$\mathbf{A} = \mathbf{K} \quad : \quad \text{prior relaxation,} \tag{C.20}$$

where $\mathbf{I}$ stands for the identity operator, $\mathbf{D}$ for a diagonal matrix, e.g., the diagonal part of $\mathbf{K}$, and $\mathbf{T}$ for a lower triangular matrix, e.g., the lower triangular part of $\mathbf{K}$. In case $\mathbf{K}$ represents the operator of the prior term in an error functional we will call iteration with $\mathbf{K}^{-1}$ (corresponding to the covariance of the prior process) *prior relaxation.*[1] For ϕ–independent $\mathbf{K}$ and T, $\eta = 1$ with invertible $\mathbf{K}$ the corresponding linear equation is solved by prior relaxation in one step. However, also linear equations have to be solved by iteration if $\mathbf{K}$ is too large to be inverted. The gradient algorithm does not

[1] For a discussion of the relation to the EM (Expectation-Maximization)-Algorithm [14, 100,104,211,350,416,477,534] see Section 9.7 of [291].

require inversion. Due to the presence of δ–peaks in the data terms, a simple gradient algorithm however does not usually work well for general density estimation problems [291].

On one hand, density estimation is a rather general problem requiring the solution of constraint, inhomogeneous, nonlinear (integro–)differential equations. On the other hand, nonparametric density estimation problems are typically 'nearly' linear — at least for Gaussian specific priors and for a simple, local relation between the states or fields $\phi(x, y)$ and the corresponding likelihoods $p(x|y, \phi)$ — and have only a relatively simple non–negativity and normalization constraint. (Due to its quantum mechanical likelihood model, the situation is more involved for inverse quantum theory.) Furthermore, the inhomogeneities are commonly restricted to a finite number of discrete training data points. Thus we expect the inversion of $\mathbf{K}$ to be the essential part of the solution for density estimation problems. However, $\mathbf{K}$ is not necessarily invertible or $\mathbf{K}^{-1}$ may be difficult to calculate. Also, inversion of $\mathbf{K}$ is not exactly what is optimal and there are improved methods. In the following, we will discuss some easy–to–implement optimization methods adapted especially to the situation of nonparametric density estimation. In particular we will recover the classical kernel methods of density estimation [456] as the (first order) approximate solution of the stationarity equations.

Clearly, if calculationally feasible, more elaborate methods like conjugate gradient or quasi–Newton approaches can also be used. On the other hand, if calculating gradient information is too costly, a possible alternative approach is the downhill simplex method [410]. Also the principles of parallel optimization methods like genetic algorithms [99, 167, 218, 274, 339, 346, 444], and of homotopy methods like simulated annealing [1, 98, 149, 261, 338, 415, 428, 445, 532–534], can be utilized.

C.2.2 Linearization and Newton algorithm

For linear equations $\mathbf{K}\phi = T$, with T and $\mathbf{K}$ not depending on ϕ, a spectral radius $\rho(\mathbf{M}) < 1$ (the largest modulus of the eigenvalues) of the *iteration matrix*

$$\mathbf{M} = -\eta \mathbf{A}^{-1}\mathbf{B}_\eta = (1-\eta)\mathbf{I} - \eta \mathbf{A}^{-1}\mathbf{B} = \mathbf{I} - \eta \mathbf{A}^{-1}\mathbf{K} \tag{C.21}$$

would guarantee convergence of the iteration scheme (C.7). This is easily seen by iterating Eq. (C.7)

$$\begin{aligned} \phi^{(i+1)} &= \eta \mathbf{A}^{-1} T + \mathbf{M}\phi^{(i)} && \text{(C.22)} \\ &= \eta \mathbf{A}^{-1} T + \eta \mathbf{M}\mathbf{A}^{-1} T + \mathbf{M}^2 \phi^{(i-1)} && \text{(C.23)} \\ &= \eta \sum_{n=0}^{\infty} \mathbf{M}^n \mathbf{A}^{-1} T. && \text{(C.24)} \end{aligned}$$

A zero mode of $\mathbf{K}$, for example a constant function for differential operators without boundary conditions, corresponds to an eigenvalue 1 of $\mathbf{M}$ and would lead to divergence of the sequence $\phi^{(i)}$.

For nonlinear equations a local convergence analysis can be based on a linear approximation around a fixed point. Expanding the gradient at some ϕ_0

$$G(\phi) = \left.\frac{\delta(-E(\phi))}{\delta\phi}\right|_{\phi_0} + (\phi - \phi_0)\,\mathbf{H}(\phi_0) + \cdots \tag{C.25}$$

shows that the factor of the linear term is the Hessian. Setting Eq. (C.25) equal to zero, and choosing a learning matrix $\mathbf{A}$ to solve $G = 0$ iteratively, the rate of convergence in the vicinity of ϕ_0 is according to Eqs. (C.21) and (C.24) determined by the spectral radius of the iteration matrix

$$\mathbf{M} = \mathbf{I} + \eta\mathbf{A}^{-1}\mathbf{H}. \tag{C.26}$$

The Newton algorithm uses the negative Hessian $-\mathbf{H}$ as learning matrix $\mathbf{A}$, provided it exists and is positive definite. Otherwise one must resort to other methods. In the linear approximation (corresponding to a quadratic energy) the Newton algorithm

$$\mathbf{A} = -\mathbf{H} \quad : \quad \text{Newton} \tag{C.27}$$

gives the solution after one iteration step and is thus optimal. In Sections 3.1.3 and 3.2.3 we have seen that the inhomogeneities in the MAP–equations of density estimation lead to an additional diagonal part in the Hessian, which can remove the zero modes of the inverse covariance $\mathbf{K}$ of a Gaussian prior factor for the Hessian.

C.2.3 Massive relaxation

We now consider methods to construct a positive definite or at least invertible learning matrix. For example, far from a minimum the Hessian $\mathbf{H}$ may not be positive definite and, like a differential operator $\mathbf{K}$ with zero modes, not even invertible. Massive relaxation can transform a non–invertible or not positive definite operator $\mathbf{A}_0$, e.g., $\mathbf{A}_0 = \mathbf{K}$ or $\mathbf{A}_0 = -\mathbf{H}$, into an invertible or positive definite operator:

$$\mathbf{A} = \mathbf{A}_0 + m^2\mathbf{I} \quad : \quad \text{massive relaxation.} \tag{C.28}$$

A generalization would be to allow $m = m(x, y)$. This is, for example, used in some realizations of Newton's method for minimization in regions where $\mathbf{H}$ is not positive definite and a diagonal operator is added to $-\mathbf{H}$, using, for example, a modified Cholesky factorization [40]. The mass term removes

the zero modes of $\mathbf{K}$ if $-m^2$ is not in the spectrum of $\mathbf{A}_0$, and makes it positive definite if m^2 is larger than the smallest eigenvalue of $\mathbf{A}_0$. Matrix elements $(\phi, (\mathbf{A}_0 - z\mathbf{I})^{-1}\phi)$ of the resolvent $\mathbf{A}^{-1}(z)$, $z = -m^2$, representing in this case a complex variable, have poles at discrete eigenvalues of $\mathbf{A}_0$ and a cut along the continuous spectrum as long as ϕ has a non–zero overlap with the corresponding eigenfunctions. Instead of multiples of the identity, other operators may also be added to remove zero modes. The Hessian $\mathbf{H}_L$ in (3.42), for example, adds an x–dependent mass–like, but not necessarily positive definite term to $\mathbf{K}$. Similarly, for example, $\mathbf{H}_P$ in (3.64) has (x,y)–dependent mass $\mathbf{P}^{-2}\mathbf{N}$ restricted to data points.

It is clear from Eq. (C.28) that prior relaxation (C.20) is the massless limit $m^2 \to 0$ of massive relaxation with $\mathbf{A}_0 = \mathbf{K}$, and a gradient algorithm (C.17) with η' can formally be obtained as infinite mass limit $m^2 \to \infty$ with $\eta \to \infty$ and $m^2/\eta = 1/\eta'$.

Constant functions are typical zero modes for differential operators with periodic boundary conditions. For instance, for a common smoothness term $-\Delta = -(\Delta_x + \Delta_y) = -\sum_{i=1}^{d_X} \partial_{x,i}^2 - \sum_{i=1}^{d_Y} \partial_{y,i}^2$ (kinetic energy operator) as regularizing operator $\mathbf{K}$ in $x = \{x_1, \cdots, x_{d_X}\}$ and $y = \{y_1, \cdots, y_{d_Y}\}$, the inverse of $\mathbf{A} = \mathbf{K} + m^2\mathbf{I}$ has the form

$$\mathbf{A}^{-1}(x', y'; x, y) = \frac{1}{-\Delta + m^2} \tag{C.29}$$

$$= \int_{-\infty}^{\infty} \frac{d^{d_X}k_x\, d^{d_Y}k_y}{(2\pi)^d} \frac{e^{ik_x(x-x')+ik_y(y-y')}}{k_x^2 + k_y^2 + m^2}, \tag{C.30}$$

with $d = d_X + d_Y$, $d_X = \dim(X)$, $d_Y = \dim(Y)$, $k_x^2 = \sum_{i=1}^{d_X} k_{x,i}^2$, and $k_y^2 = \sum_{i=1}^{d_Y} k_{y,i}^2$. This Green's function or matrix element of the resolvent kernel for $\mathbf{A}$ is analogous to the (Euclidean) propagator of a free scalar field with mass m, which is its two–point correlation function or matrix element of the covariance operator. According to $1/x = \int_0^\infty dt\, e^{-xt}$ the denominator can be brought into the exponent by introducing an additional integral. Performing the resulting Gaussian integration over $k = (k_x, k_y)$ the inverse can be expressed as

$$\mathbf{A}^{-1}(x', y'; x, y; m) = m^{d-2}\mathbf{A}^{-1}(m(x - x'), m(y - y'); 1)$$

$$= (2\pi)^{-d/2} \left(\frac{m}{|x - x'| + |y - y'|}\right)^{(d-2)/2} K_{(d-2)/2}(m|x-x'|+m|y-y'|) \tag{C.31}$$

in terms of the modified Bessel functions $K_\nu(x)$ which have the following integral representation

$$K_\nu(2\sqrt{\beta\gamma}) = \frac{1}{2}\left(\frac{\gamma}{\beta}\right)^{\frac{\nu}{2}} \int_0^\infty dt\, t^{\nu-1} e^{\frac{\beta}{t} - \gamma t}. \tag{C.32}$$

Alternatively, the same result can be obtained by switching to d–dimensional spherical coordinates, expanding the exponential in ultra-spheric harmonic

functions and performing the integration over the angle-variables [264]. For the example $d = 1$ this corresponds to Parzen's kernel used in density estimation or for $d = 3$

$$\mathbf{A}^{-1}(x', y'; x, y) = \frac{1}{4\pi|x - x'| + 4\pi|y - y'|} e^{-m|x-x'|-m|y-y'|}. \qquad \text{(C.33)}$$

The Green's function for periodic, Neumann, or Dirichlet boundary conditions can be expressed by sums over $\mathbf{A}^{-1}(x', y'; x, y)$ [165].

The lattice version of the Laplacian with lattice spacing a reads

$$\hat{\Delta} f(n) = \frac{1}{a^2} \sum_{j}^{d} [f(n - a_j) - 2f(n) + f(n + a_j)], \qquad \text{(C.34)}$$

writing a_j for a vector in direction j and length a. Including a mass term we get as lattice approximation for $\mathbf{A}$

$$\hat{\mathbf{A}}(n_x, n_y; m_x, m_y) = -\frac{1}{a^2} \sum_{i=1}^{d_X} \delta_{n_y,m_y} (\delta_{n_x+a_i^x,m_x} - 2\delta_{n_x,m_x} + \delta_{n_x-a_i^x,m_x})$$

$$-\frac{1}{a^2} \sum_{j=1}^{d_Y} \delta_{n_x,m_x} (\delta_{n_y+a_j^y,m_y} - 2\delta_{n_y,m_y} + \delta_{n_y-a_j^y,m_y}) + m^2 \delta_{n_x,m_x} \delta_{n_y,m_y}. \qquad \text{(C.35)}$$

Inserting the Fourier representation of the Kronecker symbol $\delta_{n,m}$, i.e., the lattice version of (2.118), gives

$$\hat{\mathbf{A}}(n_x, n_y; m_x, m_y) = \frac{2d}{a^2} \int_{-\pi}^{\pi} \frac{d^{d_X}k_x \, d^{d_Y}k_y}{(2\pi)^d} e^{ik_x(n_x-m_x)+ik_y(n_y-m_y)}$$

$$\times \left(1 + \frac{m^2a^2}{2d} - \frac{1}{d} \sum_{i=1}^{d_X} \cos k_{x,i} - \frac{1}{d} \sum_{j=1}^{d_Y} \cos k_{y,j} \right), \qquad \text{(C.36)}$$

with $k_{x,i} = k_x a_i^x$, $\cos k_{y,j} = \cos k_y a_j^y$ and inverse

$$\hat{\mathbf{A}}^{-1}(n_x, n_y; m_x, m_y) = \int_{-\pi}^{\pi} \frac{d^{d_X}k_x \, d^{d_Y}k_y}{(2\pi)^d} \hat{\mathbf{A}}^{-1}(k_x, k_y) e^{ik_x(n_x-m_x)+ik_y(n_y-m_y)}$$

$$= \frac{a^2}{2d} \int_{-\pi}^{\pi} \frac{d^{d_X}k_x \, d^{d_Y}k_y}{(2\pi)^d} \frac{e^{ik_x(n_x-m_x)+ik_y(n_y-m_y)}}{1 + \frac{m^2a^2}{2d} - \frac{1}{d}\sum_{i=1}^{d_X} \cos k_{x,i} - \frac{1}{d}\sum_{j=1}^{d_Y} \cos k_{y,j}}. \qquad \text{(C.37)}$$

For $m = 0$ and $d \leq 2$ the integrand diverges for $k \to 0$ ('infrared divergence'). Subtracting formally the infinite $\hat{\mathbf{A}}^{-1}(0, 0; 0, 0)$ results in a finite difference.

For example, in $d = 1$ one finds $\hat{\mathbf{A}}^{-1}(n_y; m_y) - \hat{\mathbf{A}}^{-1}(0;0) = -(1/2)|n_y - m_y|$ [232]. Using again $1/x = \int_0^\infty dt\, e^{-xt}$ one obtains

$$\hat{\mathbf{A}}^{-1}(k_x, k_y) = \frac{1}{2}\int_0^\infty dt\, e^{-\mu t + a^{-2}t\left(\sum_i^{d_X}\cos k_{x,i} + \sum_j^{d_Y}\cos k_{y,j}\right)}, \tag{C.38}$$

with $\mu = d/a^2 + m^2/2$. This allows to express the inverse $\hat{\mathbf{A}}^{-1}(k_x, k_y)$ in terms of the modified Bessel functions $I_\nu(n)$ which have for integer argument n the integral representation

$$I_\nu(n) = \frac{1}{\pi}\int_0^\pi d\Theta\, e^{n\cos\Theta}\cos(\nu\Theta). \tag{C.39}$$

One finds

$$\hat{\mathbf{A}}^{-1}(n_x, n_y; m_x, m_y) =$$

$$\frac{1}{2}\int_0^\infty e^{-\mu t}\prod_{i=1}^{d_X} K_{|n_{x,i}-n'_{x,i}|}(t/a^2)\prod_{j=1}^{d_Y} K_{|m_{y,j}-m'_{y,j}|}(t/a^2). \tag{C.40}$$

It might be interesting to remark that the matrix elements of the inverse learning matrix or free massive propagator on the lattice $\hat{\mathbf{A}}^{-1}(x', y'; x, y)$ can be given an interpretation in terms of (random) walks connecting the two points (x', y') and (x, y) [125,408]. For that purpose the lattice Laplacian is split into a diagonal and a nearest neighbor part

$$-\hat{\Delta} = \frac{1}{a^2}\left(2d\mathbf{I} - \mathbf{W}\right), \tag{C.41}$$

where the nearest neighbor matrix $\mathbf{W}$ has matrix elements equal to one for nearest neighbors and equal to zero otherwise. Thus

$$\left(-\hat{\Delta} + m^2\right)^{-1} = \frac{1}{2\mu}\left(\mathbf{I} - \frac{1}{2\mu a^2}\mathbf{W}\right)^{-1} = \frac{1}{2\mu}\sum_{n=0}^{\infty}\left(\frac{1}{2\mu a^2}\right)^n \mathbf{W}^n \tag{C.42}$$

can be written as geometric series. The matrix elements $\mathbf{W}^n(x', y'; x, y)$ give the number of walks $w[(x', y') \to (x, y)]$ of length $|w| = n$ connecting the two points (x', y') and (x, y). Thus one can write

$$\left(-\hat{\Delta} + m^2\right)^{-1}(x', y'; x, y) = \frac{1}{2\mu}\sum_{w[(x',y')\to(x,y)]}\left(\frac{1}{2\mu a^2}\right)^{|w|}. \tag{C.43}$$

C.2.4 Gaussian relaxation

As Gaussian kernels are often used in density estimation and also in function approximation (e.g., for Radial Basis Functions [124, 399, 400, 507, 510]) we consider the example

$$\mathbf{A} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\mathbf{M}^2}{2\tilde{\sigma}^2} \right)^k = e^{\frac{\mathbf{M}^2}{2\tilde{\sigma}^2}} \quad : \text{ (inverse) Gaussian,} \tag{C.44}$$

with some positive semi–definite $\mathbf{M}^2$. The contribution for $k = 0$ has the form of a mass term, so for positive semi–definite $\mathbf{M}$ this $\mathbf{A}$ is positive definite and therefore invertible with inverse

$$\mathbf{A}^{-1} = e^{-\frac{\mathbf{M}^2}{2\tilde{\sigma}^2}}, \tag{C.45}$$

which is diagonal and Gaussian in $\mathbf{M}$–representation. In the limit $\tilde{\sigma} \to \infty$ or for zero modes of $\mathbf{M}$, the Gaussian $\mathbf{A}^{-1}$ becomes the identity $\mathbf{I}$, corresponding to the gradient algorithm. Consider

$$\mathbf{M}^2(x', y'; x, y) = -\delta(x - x')\delta(y - y')\Delta \tag{C.46}$$

where the δ–functions are usually skipped from the notation and $\Delta = \Delta_x + \Delta_y = \sum_{i=1}^{d_X} \partial_{x,i}^2 + \sum_{i=1}^{d_Y} \partial_{y,i}^2$ denotes the Laplacian. The kernel of the inverse is diagonal in Fourier representation

$$A^{(-1)}(k_x', k_y';, k_x, k_y) = \delta(k_x - k_x')\delta(k_y - k_y')e^{-\frac{k_x^2 + k_y^2}{2\tilde{\sigma}^2}}, \tag{C.47}$$

whereas non–diagonal but also Gaussian in (x, y)–representation

$$\mathbf{A}^{-1}(x', y'; x, y) = e^{-\frac{\Delta}{2\tilde{\sigma}^2}} = \int \frac{dk_x dk_y}{(2\pi)^d} e^{-\frac{k_x^2 + k_y^2}{2\tilde{\sigma}^2} + ik_x(x - x') + ik_y(y - y')} \tag{C.48}$$

$$= \left(\frac{\tilde{\sigma}}{\sqrt{2\pi}} \right)^d e^{-\tilde{\sigma}^2((x-x')^2 + (y-y')^2)} = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^d} e^{-\frac{(x-x')^2 + (y-y')^2}{2\sigma^2}}, \tag{C.49}$$

with $\sigma = 1/\tilde{\sigma}$ and $d = d_X + d_Y$, $d_X = \dim(X)$, $d_Y = \dim(Y)$.

C.2.5 Inverting in subspaces

Matrices considered as learning matrix have to be invertible. Non-invertible matrices can only be inverted in the complement of its null space. With respect to a symmetric $\mathbf{A}$ we define the projector $\mathbf{Q}_0 = \mathbf{I} - \sum_i \psi_i \psi_i^T$ into its null space (for the more general case of a hermitian $\mathbf{A}$ replace ψ_i^T by the hermitian conjugate $\psi_i^\dagger$) and its complement $\mathbf{Q}_1 = \mathbf{I} - \mathbf{Q}_0 = \sum_i \psi_i \psi_i^T$ with ψ_i denoting orthogonal eigenvectors with eigenvalues $a_i \neq 0$ of $\mathbf{A}$, i.e.,

$\mathbf{A}\psi_i = a_i\psi_i \neq 0$. Then, denoting projected sub-matrices by $\mathbf{Q}_i\mathbf{A}\mathbf{Q}_j = \mathbf{A}_{ij}$ we have $\mathbf{A}_{00} = \mathbf{A}_{10} = \mathbf{A}_{01} = 0$, and $\mathbf{A}$ reduces to

$$\mathbf{A} = \mathbf{Q}_1\mathbf{A}\mathbf{Q}_1 = \mathbf{A}_{11}. \tag{C.50}$$

In the update equation (C.10)

$$\mathbf{A}\Delta\phi^{(i)} = \eta G \tag{C.51}$$

only $\mathbf{A}_{11}$ can be inverted. Writing $\mathbf{Q}_j\phi = \phi_j$, $j = 0,1$, for a projected vector, and analogously $\mathbf{Q}_i G = G_i$, the iteration scheme (C.10) acquires the form

$$\Delta\phi_1^{(i)} = \eta\mathbf{A}_{11}^{-1}G_1, \tag{C.52}$$

$$0 = \eta G_0. \tag{C.53}$$

For positive semi–definite $\mathbf{A}$ the sub-matrix $\mathbf{A}_{11}$ is positive definite. If the second equation is already fulfilled or its solution is postponed to a later iteration step, we have

$$\phi_1^{(i+1)} = \phi_1^{(i)} + \eta\mathbf{A}_{11}^{-1}\left(T_1^{(i)} - \mathbf{K}_{11}^{(i)}\phi_1^{(i)} - \mathbf{K}_{10}^{(i)}\phi_0^{(i)}\right), \tag{C.54}$$

$$\phi_0^{(i+1)} = \phi_0^{(i)}. \tag{C.55}$$

In case the projector $\mathbf{Q}_0 = \mathbf{I}_0$ is diagonal in the chosen representation, the projected equation can directly be solved by skipping the corresponding components. Otherwise one can use the Moore–Penrose inverse $\mathbf{A}^{\#}$ of $\mathbf{A}$ to solve the projected equation

$$\Delta\phi^{(i)} = \eta\mathbf{A}^{\#}G. \tag{C.56}$$

Alternatively, an invertible operator $\tilde{\mathbf{A}}_{00}$ can be added to $\mathbf{A}_{11}$ to obtain a complete iteration scheme with $\mathbf{A}^{-1} = \mathbf{A}_{11}^{-1} + \tilde{\mathbf{A}}_{00}^{-1}$

$$\begin{aligned}\phi^{(i+1)} = {} & \phi^{(i)} + \eta\mathbf{A}_{11}^{-1}\left(T_1^{(i)} - \mathbf{K}_{11}^{(i)}\phi_1^{(i)} - \mathbf{K}_{10}^{(i)}\phi_0^{(i)}\right) \\ & + \eta\tilde{\mathbf{A}}_{00}^{-1}\left(T_0^{(i)} - \mathbf{K}_{01}^{(i)}\phi_1^{(i)} - \mathbf{K}_{00}^{(i)}\phi_0^{(i)}\right).\end{aligned} \tag{C.57}$$

The choice $\mathbf{A}^{-1} = (\mathbf{A}_{11} + \mathbf{I}_{00})^{-1} = \mathbf{A}_{11}^{-1} + \mathbf{I}_{00} = \mathbf{A}_{11}^{-1} + \mathbf{Q}_0$, for instance, results in a gradient algorithm on the null space with additional coupling between the two subspaces.

C.2.6 Boundary conditions

For a differential operator invertibility can be achieved by adding an operator restricted to a subset $B \subset X \times Y$ (boundary). In a more general example, we consider a projector $\mathbf{Q}_B$ on a space which we will call boundary and the projector on the interior $\mathbf{Q}_I = \mathbf{I} - \mathbf{Q}_B$. We write $\mathbf{Q}_k\mathbf{K}\mathbf{Q}_l = \mathbf{K}_{kl}$ for

$k, l \in \{I, B\}$, and require $\mathbf{K}_{BI} = 0$. That means $\mathbf{K}$ is not symmetric, but $\mathbf{K}_{II}$ can be, and we have

$$\mathbf{K} = \mathbf{K}_{II} + \mathbf{K}_{IB} + \mathbf{K}_{BB}. \tag{C.58}$$

For such $\mathbf{K}$, an equation of the form $\mathbf{K}\phi = T$ can be decomposed into

$$\mathbf{K}_{BB}\phi_B = T_B, \tag{C.59}$$

$$\mathbf{K}_{IB}\phi_B + \mathbf{K}_{II}\phi_I = T_I, \tag{C.60}$$

with projected $\phi_k = \mathbf{Q}_k\phi$, $T_k = \mathbf{Q}_k T$, so that

$$\phi_B = \mathbf{K}_{BB}^{-1} T_B, \tag{C.61}$$

$$\phi_I = \mathbf{K}_{II}^{-1}\left(T_I - \mathbf{K}_{IB}\mathbf{K}_{BB}^{-1} T_B\right). \tag{C.62}$$

The boundary part is independent of the interior, however, the interior can depend on the boundary. A basis can be chosen so that the projector onto the boundary is diagonal,

$$\mathbf{Q}_B = \mathbf{I}_B = \sum_{j:(x_j,y_j)\in B} (\delta(x_j) \otimes \delta(y_j)) \otimes (\delta(x_j) \otimes \delta(y_j))^T.$$

Eliminating the boundary results in an equation for the interior with adapted inhomogeneity. The special case $\mathbf{K}_{BB} = \mathbf{I}_{BB}$, i.e., $\phi_B = T_B$ on the boundary, is known as Dirichlet boundary condition.

As a trivial example of an equation $\mathbf{K}\phi = T$ with boundary conditions, consider a one–dimensional finite difference approximation for a negative Laplacian $\mathbf{K}$, adapted as in Eq. (C.58) to include the boundary conditions $\phi_1 = b$, $\phi_6 = b$,

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \end{pmatrix} = \begin{pmatrix} b \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ b \end{pmatrix}. \tag{C.63}$$

Then Eq. (C.59) is equivalent to the two boundary conditions, $\phi_1 = b$, $\phi_6 = b$, and the interior equation Eq. (C.60) reads

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{pmatrix} = \begin{pmatrix} T_2 \\ T_3 \\ T_4 \\ T_5 \end{pmatrix} + \begin{pmatrix} b \\ 0 \\ 0 \\ b \end{pmatrix}. \tag{C.64}$$

Useful references dealing with the numerical solution of partial differential equations are, for example, [17, 184, 189, 345, 410].

Similar to boundary conditions for $\mathbf{K}$, we may use a learning matrix $\mathbf{A}$ with boundary conditions (corresponding, for example, to those used for $\mathbf{K}$):

$$\mathbf{A} = \mathbf{A}_{II} + \mathbf{A}_{IB} + \mathbf{A}_{BB} \quad : \quad \text{Boundary} \tag{C.65}$$

$$\mathbf{A} = \mathbf{A}_{II} + \mathbf{A}_{IB} + \mathbf{I}_{BB} \quad : \quad \text{Dirichlet boundary.} \tag{C.66}$$

For linear $\mathbf{A}_{BB}$ the form (C.65) corresponds to general linear boundary conditions. (It is also possible to require nonlinear boundary conditions.) $\mathbf{A}_{II}$ can be chosen symmetric, and therefore positive definite, and the boundary of $\mathbf{A}$ can be changed during iteration. Solving $\mathbf{A}(\phi^{(i+1)} - \phi^{(i)}) = \eta(T^{(i)} - \mathbf{K}^{(i)}\phi^{(i)})$ gives on the boundary and for the interior

$$\phi_B^{(i+1)} = \phi_B^i + \eta \mathbf{A}_{BB}^{-1} \left(T_B^{(i)} - \mathbf{K}_{BB}^{(i)} \phi_B^{(i)} - \mathbf{K}_{BI}^{(i)} \phi_I^{(i)} \right), \tag{C.67}$$

$$\phi_I^{(i+1)} = \phi_I^i + \eta \mathbf{A}_{II}^{-1} \left(T_I^{(i)} - \mathbf{K}_{II}^{(i)} \phi_I^{(i)} - \mathbf{K}_{IB}^{(i)} \phi_B^{(i)} \right) - \mathbf{A}_{II}^{-1} \mathbf{A}_{IB} \left(\phi_B^{(i+1)} - \phi_B^{(i)} \right). \tag{C.68}$$

For fulfilled boundary conditions with $\phi_B^{(i)} = \left(\mathbf{K}_{BB}^{(i)} \right)^{-1} T_B^{(i)}$ and $\mathbf{K}_{BI}^{(i)} = 0$, or for $\eta \mathbf{A}_{BB}^{-1} \to 0$ so that the boundary is not updated, the term $\phi_B^{(i+1)} - \phi_B^{(i)}$ vanishes. Otherwise, inserting the first in the second equation gives

$$\begin{aligned} \phi_I^{(i+1)} \quad = \quad & \phi_I^i + \eta \mathbf{A}_{II}^{-1} \left(T_I^{(i)} - \mathbf{K}_{II}^{(i)} \phi_I^{(i)} - \mathbf{K}_{IB}^{(i)} \phi_B^{(i)} \right) \\ & - \eta \mathbf{A}_{II}^{-1} \mathbf{A}_{IB} \mathbf{A}_{BB}^{-1} \left(T_B^{(i)} - \mathbf{K}_{BB}^{(i)} \phi_B^{(i)} - \mathbf{K}_{BI}^{(i)} \phi_I^{(i)} \right). \end{aligned} \tag{C.69}$$

Even if $\mathbf{K}$ is not defined with boundary conditions, an invertible $\mathbf{A}$ can be obtained from $\mathbf{K}$ by introducing a boundary for $\mathbf{A}$. The updating process is then restricted to the interior. In such cases the boundary should be systematically changed during iteration. Block–wise updating of ϕ represents a special case of such learning matrices with variable boundary.

Table C.2 summarizes the learning matrices we have discussed in some detail for density estimation (for conjugate gradient and quasi–Newton methods see, for example, [410]).

C.3 Initial configurations and kernel methods

C.3.1 Truncated equations

To solve the nonlinear Eq. (C.3) by iteration one has to begin with an initial configuration $\phi^{(0)}$. In principle any easy to use technique for density estimation could be chosen to construct starting guesses $\phi^{(0)}$. One possibility to obtain initial guesses is to neglect some terms of the full stationarity equation and solve the resulting simpler (ideally linear) equation first. The

Learning algorithm	Learning matrix
Gradient	$\mathbf{A} = \mathbf{I}$
Jacobi	$\mathbf{A} = \mathbf{D}$
Gauss–Seidel	$\mathbf{A} = \mathbf{T} + \mathbf{D}$
Newton	$\mathbf{A} = -\mathbf{H}$
prior relaxation	$\mathbf{A} = \mathbf{K}$
massive relaxation	$\mathbf{A} = \mathbf{A}_0 + m^2\mathbf{I}$
linear boundary	$\mathbf{A} = \mathbf{A}_{II} + \mathbf{A}_{IB} + \mathbf{A}_{BB}$
Dirichlet boundary	$\mathbf{A} = \mathbf{A}_{II} + \mathbf{A}_{IB} + \mathbf{I}_{BB}$
(inverse) Gaussian	$\mathbf{A} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\mathbf{M}^2}{2\delta^2}\right)^k = e^{\frac{\mathbf{M}^2}{2\delta^2}}$

Table C.1: Learning matrices.

corresponding solution may be taken as initial guess $\phi^{(0)}$ for solving the full equation.

Typical error functionals for statistical learning problems, like (4.4), include a term $\langle L|N\rangle$ consisting of a discrete sum over a finite number n of training data. For diagonal $\mathbf{P}'$ those contributions result in n δ–peak contributions to the inhomogeneities T of the stationarity equations, like $\sum_i \delta(x-x_i)\delta(y-y_i)$ in Eq. (3.28) or $\sum_i \delta(x-x_i)\delta(y-y_i)/P(x,y)$ in Eq. (3.55). To find an initial guess, one can now keep only those δ–peak contributions T_δ arising from the training data and ignore the other, typically continuous parts of T. For (3.28) and (3.55) this means setting $\Lambda_X = 0$ and yields a truncated equation

$$\mathbf{K}\phi = \mathbf{P}'\mathbf{P}^{-1}N = T_\delta. \tag{C.70}$$

Hence ϕ can, for diagonal $\mathbf{P}'$, be written as a sum of n terms

$$\phi(x,y) = \sum_{i=1}^{n} \mathbf{C}(x,y;x_i,y_i)\frac{P'(x_i,y_i)}{P(x_i,y_i)}, \tag{C.71}$$

with $\mathbf{C} = \mathbf{K}^{-1}$, provided the inverse $\mathbf{K}^{-1}$ exists. For $E(L)$ the resulting truncated equation is linear in L. For $E(P)$, however, the truncated equation remains nonlinear. Having solved the truncated equation we restore the necessary constraints for ϕ, like normalization and non–negativity for P or normalization of the exponential for L. (A table of common choices for kernel functions can be found in [124].) In general, some $\mathbf{C} \neq \mathbf{K}^{-1}$ can be chosen. This is necessary if $\mathbf{K}$ is not invertible and can also be useful if its inverse is difficult to calculate. Deriving the kernel from a prior process one possible choice for the kernel is the inverse negative Hessian $\mathbf{C} = -\mathbf{H}^{-1}$, evaluated

at a chosen initial configuration $\tilde{\phi}^{(0)}$, or an approximation of it. A simple possibility to construct an invertible operator from non–invertible $\mathbf{K}$ would be to add a mass term

$$\mathbf{C} = \left(\mathbf{K} + m_C^2 \mathbf{I}\right)^{-1} \tag{C.72}$$

or to impose additional boundary conditions.

Solving a truncated equation of the form (C.71) with $\mathbf{C}$ means skipping the term $-\mathbf{C}(\mathbf{P}'\Lambda_X + (\mathbf{K} - \mathbf{C}^{-1})\phi)$ from the exact relation (3.101)

$$\phi = \mathbf{C}\mathbf{P}'\mathbf{P}^{-1}N - \mathbf{C}(\mathbf{P}'\Lambda_X + (\mathbf{K} - \mathbf{C}^{-1})\phi). \tag{C.73}$$

A kernel used to create an initial guess $\phi^{(0)}$ will be called an *initializing kernel.*

A similar possibility is to start with an 'empirical solution'

$$\phi^{(0)} = \phi_{\text{emp}}, \tag{C.74}$$

where ϕ_{emp} is defined as a ϕ which reproduces the conditional empirical density P_{emp} of Eq. (3.71) obtained from the training data,

$$P_{\text{emp}} = P(\phi_{\text{emp}}). \tag{C.75}$$

In case there are not data points for every x–value, a correctly normalized initial solution would, for example, be given by $\tilde{P}_{\text{emp}}$ defined in Eq. (3.72). If zero values of the empirical density correspond to infinite values for ϕ, like in the case $\phi = L$, one can use $P^{\epsilon}_{\text{emp}}$ as defined in Eq. (3.73), with small ϵ, to obtain an initial guess.

Similar to Eq. (C.71), it is often also useful to choose a (for example, smoothing) kernel $\mathbf{C}$ and use as initial guess

$$\phi^{(0)} = \mathbf{C}\phi_{\text{emp}}, \tag{C.76}$$

or a properly normalized version thereof. Alternatively one may also let the (smoothing) operator $\mathbf{C}$ act directly on P_{emp} and use a corresponding ϕ as initial guess,

$$\phi^{(0)} = \phi(\mathbf{C}P_{\text{emp}}), \tag{C.77}$$

assuming an inverse $\phi(P)$ of the mapping $P(\phi)$ exists.

We will now discuss the cases $\phi = L$ and $\phi = P$ in some more detail.

C.3.2 Kernels for L

For functional $E(L)$ a truncated equation is of the form

$$L = \mathbf{C}N. \tag{C.78}$$

For the choice $\phi = L$ we have $\mathbf{P}'\mathbf{P}^{-1} = \mathbf{I}$, so Eq. (C.70) would be obtained for $\mathbf{C} = \mathbf{K}^{-1}$. As $\mathbf{K}^{-1}$ does not necessarily exist, may be difficult to calculate, or

one is looking for a normalized guess $L^{(0)}$, $\mathbf{C}$ are also considered, which are not equal to $\mathbf{K}^{-1}$ but still keep some essential features of it.

Normalizing the exponential of the solution gives

$$L(x,y) = \sum_i^n \mathbf{C}(x,y;x_i,y_i) - \ln \int dy' \, e^{\sum_i^n \mathbf{C}(x,y';x_i,y_i)}, \tag{C.79}$$

or

$$L = \mathbf{C}N - \ln \mathbf{I}_X e^{\mathbf{C}N}. \tag{C.80}$$

Notice that when normalizing L according to Eq. (C.79) after each iteration, the truncated equation (C.78) is equivalent to a one–step iteration with uniform $P^{(0)} = e^{L^{(0)}}$ according to

$$L^1 = \mathbf{C}N + \mathbf{C}\mathbf{P}^{(0)}\Lambda_X, \tag{C.81}$$

where only $(\mathbf{I} - \mathbf{C}\mathbf{K})L$ is missing from the nontruncated equation (C.73), because the additional y–independent term $\mathbf{C}\mathbf{P}^{(0)}\Lambda_X$ becomes inessential if L is normalized afterwards.

Let us consider as an example the choice $\mathbf{C} = -\mathbf{H}^{-1}(\phi^{(0)})$ for uniform initial $L^{(0)} = c$, corresponding to a normalized P and $\mathbf{K}L^{(0)} = 0$ (e.g., for a differential operator $\mathbf{K}$). Uniform $L^{(0)}$ means uniform $P^{(0)} = 1/v_y$, assuming that $v_y = \int dy$ exists and, according to Eq. (3.23), $\Lambda_X = N_X$ for $\mathbf{K}L^{(0)} = 0$. Thus the Hessian (3.45) at $L^{(0)}$ is found as

$$\mathbf{H}(L^{(0)}) = -\left(\mathbf{I} - \frac{\mathbf{I}_X}{v_y}\right)\mathbf{K}\left(\mathbf{I} - \frac{\mathbf{I}_X}{v_y}\right) - \left(\mathbf{I} - \frac{\mathbf{I}_X}{v_y}\right)\frac{\mathbf{N}_X}{v_y} = -\mathbf{C}^{-1}, \tag{C.82}$$

which can be invertible due to the presence of the second term.

Another possibility is to start with an approximate empirical log–density, defined as

$$L^{\epsilon}_{\text{emp}} = \ln P^{\epsilon}_{\text{emp}}, \tag{C.83}$$

with $P^{\epsilon}_{\text{emp}}$ given in Eq. (3.73). Analogous to Eq. (C.76), the empirical log–density may, for example, also be smoothed and correctly normalized again, resulting in an initial guess,

$$L^{(0)} = \mathbf{C}L^{\epsilon}_{\text{emp}} - \ln \mathbf{I}_X e^{\mathbf{C}L^{\epsilon}_{\text{emp}}}. \tag{C.84}$$

Similarly, one may let a kernel $\mathbf{C}$, or its normalized version $\tilde{\mathbf{C}}$ defined below in Eqs. (C.88) and (C.89), act on P_{emp} first and then take the logarithm

$$L^{(0)} = \ln(\tilde{\mathbf{C}}P^{\epsilon}_{\text{emp}}). \tag{C.85}$$

Because $\tilde{\mathbf{C}}P_{\text{emp}}$ is typically nonzero already, it is usually not necessary to work here with $P^{\epsilon}_{\text{emp}}$. Like in the next section P_{emp} may be also be replaced by $\tilde{P}_{\text{emp}}$ as defined in Eq. (3.72).

C.3.3 Kernels for P

For $E(P)$ the truncated equation

$$P = \mathbf{C}\mathbf{P}^{-1}N \tag{C.86}$$

is still nonlinear in P. If we solve this equation approximately by a one–step iteration $P^1 = \mathbf{C}(\mathbf{P}^{(0)})^{-1}N$, starting from a uniform initial $P^{(0)}$ and normalizing afterwards, this corresponds for a single x–value to the classical kernel methods commonly used in density estimation. As normalized density results

$$P(x,y) = \frac{\sum_i \mathbf{C}(x,y;x_i,y_i)}{\int dy' \sum_i \mathbf{C}(x,y';x_i,y_i)} = \sum_i \bar{\mathbf{C}}(x,y;x_i,y_i), \tag{C.87}$$

i.e.,

$$P = \mathbf{N}_{C,X}^{-1}\mathbf{C}N = \bar{\mathbf{C}}N, \tag{C.88}$$

with (data dependent) normalized kernel $\bar{\mathbf{C}} = \mathbf{N}_{C,X}^{-1}\mathbf{C}$ and $\mathbf{N}_{C,X}$ the diagonal matrix with diagonal elements $\mathbf{I}_X\mathbf{C}N$. Again $\mathbf{C} = \mathbf{K}^{-1}$ or similar invertible choices can be used to obtain a starting guess for P. The form of the Hessian (3.64) suggests in particular to include mass terms on the data points.

It is interesting to interpret Eq. (C.88) as stationarity equation of a functional $\hat{E}(P)$ containing the usual data term $\sum_i \ln P(x_i,y_i)$. To obtain the derivative $\mathbf{P}^{-1}N$ of this data term we multiply, for existing $\bar{\mathbf{C}}^{-1}$, Eq. (C.88) by $\mathbf{P}^{-1}\bar{\mathbf{C}}^{-1}$, where $P \neq 0$ at data points, so that

$$\tilde{\mathbf{C}}^{-1}P = \mathbf{P}^{-1}N, \tag{C.89}$$

with data dependent

$$\tilde{\mathbf{C}}^{-1}(x,y;x',y') = \frac{\bar{\mathbf{C}}^{-1}(x,y;x',y')}{\sum_i \bar{\mathbf{C}}(x,y;x_i,y_i)}. \tag{C.90}$$

Thus Eq. (C.88) is the stationarity equation of a functional with (data–dependent) covariance $\tilde{\mathbf{C}}$

$$\hat{E}(P) = -\langle \ln P | N \rangle + \frac{1}{2}\left\langle P \left| \tilde{\mathbf{C}}^{-1} \right| P \right\rangle. \tag{C.91}$$

To study the dependence on the number n of training data for a given $\mathbf{C}$, consider a normalized kernel with $\int dy\, \mathbf{C}(x,y;x',y') = \lambda$, $\forall x,x',y'$. For such a kernel the denominator of $\bar{\mathbf{C}}$ is equal to $n\lambda$, so we have

$$\bar{\mathbf{C}} = \frac{\mathbf{C}}{n\lambda}, \quad P = \frac{\mathbf{C}N}{n\lambda}. \tag{C.92}$$

Assuming that for large n the empirical average $(1/n)\sum_i \mathbf{C}(x,y;x_i,y_i)$ in the denominator of $\widetilde{\mathbf{C}}^{-1}$ becomes n independent, e.g., converging to the true average $\int dx'dy'\, p(x',y')\mathbf{C}(x,y;x',y')$, the regularizing term in functional (C.91) becomes proportional to n

$$\widetilde{\mathbf{C}}^{-1} \propto n\lambda^2. \tag{C.93}$$

According to Eq. (2.97) this would allow a saddle point approximation to relate to a large n–limit.

Again, a similar possibility is to start with the empirical density $\tilde{P}_{\text{emp}}$ defined in Eq. (3.72). Analogous to Eq. (C.76), the empirical density can, for example, also be smoothed and correctly normalized again, so that

$$P^{(0)} = \tilde{\mathbf{C}}\tilde{P}_{\text{emp}}, \tag{C.94}$$

with $\tilde{\mathbf{C}}$ defined in Eq. (C.88).

Fig. C.1 compares the initialization according to Eq. (C.87), where the smoothing operator $\tilde{C}$ acts on N, with an initialization according to Eq. (C.94), where the smoothing operator $\tilde{C}$ acts on the correctly normalized $\tilde{P}_{\text{emp}}$.

Finally, let us show how the general framework for density estimation can be applied to a classification problem. In classification problems the dependent variable y takes only discrete values. Hence the integral over y in Eq. (C.87) becomes a sum over the discrete values of y,

$$P(x,y) = \frac{\sum_i \mathbf{C}(x,y,x_i,y_i)}{\sum_{y=0}^{1}\sum_i \mathbf{C}(x,y,x_i,y_i)}. \tag{C.95}$$

In classification it is usually assumed that the kernel $\mathbf{C}$ is diagonal in the class variable y so that

$$\mathbf{C}(x,y,x_i,y_i) = \delta_{y,y_i}\mathbf{C}(x,x_i). \tag{C.96}$$

The likelihoods $P(x,y) = p(y|x)$ are then only related through the normalization for each x in Eq. (C.87). The diagonal structure in Eq. (C.96), however, is not a necessary condition for classification, and in cases where some *a priori* information is available relating $p(y|x)$ for different classes, this can be implemented straightforwardly.

Figure C.2 shows an example with two classes decoded by $y = 1$ and $y = 0$. For the example a Gaussian kernel $\mathbf{C}(x,y,x_i,y_i)$ has been chosen, diagonal in the class variable y,

$$\begin{aligned}
\mathbf{C}(x,y,x_i,y_i) &= \delta_{y,y_i}\mathbf{C}(x,x_i) \qquad (C.97)\\
&= \delta_{y,y_i}\frac{e^{-\frac{(x-x_i)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\\
&= \frac{yy_i + (1-y)(1-y_i)}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right).
\end{aligned}$$

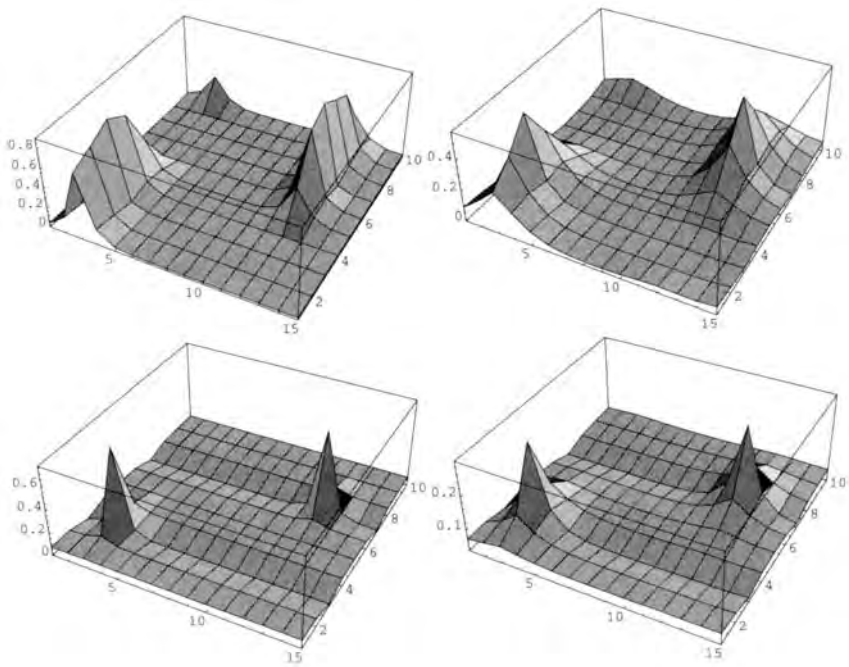


Figure C.1: Initial guesses for P. Comparison of initial guesses $P^{(0)}(x, y)$ for a case with two data points located at $(3, 3)$ and $(7, 12)$ within the intervals $y \in [1, 15]$ and $x \in [1, 10]$ with periodic boundary conditions. First row: $P^{(0)} = \tilde{\mathbf{C}}N$. (The smoothing operator acts on the unnormalized N. The following conditional normalization changes the shape more drastically than in the example shown in the second row.) Second row: $P^{(0)} = \tilde{\mathbf{C}}\tilde{P}_{\text{emp}}$. (The smoothing operator acts on the already conditionally normalized $\tilde{P}_{\text{emp}}$.) The kernel $\tilde{\mathbf{C}}$ is given by Eq. (C.88) with $\mathbf{C} = (\mathbf{K} + m_C^2\mathbf{I})$, $m_C^2 = 1.0$, and a $\mathbf{K}$ of the form of Eq. (3.83) with $\lambda_0 = \lambda_4 = \lambda_6 = 0$, and $\lambda_2 = 0.1$ (figures on the left hand side) or $\lambda_2 = 1.0$ (figures on the right hand side), respectively.

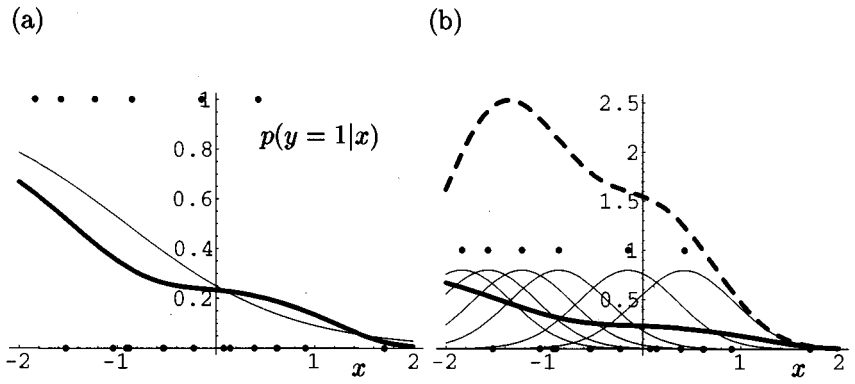


Figure C.2: Nonparametric classification: An example. Nonparametric classification with a Gaussian kernel, diagonal in y. (a) Reconstruction of the likelihood: 20 data points (x_i, y_i) (black dots) have been sampled using a true likelihood $p_{\text{true}}(y = 1|x) = \frac{1}{1+\exp(a+bx)}$ (thin line) with $a = 1.1$, $b = 1.2$, and x distributed according to $p(x) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2}\right)$ with $x_0 = 0$ and $\sigma_x = 1$. The reconstructed likelihood (thick line) is given by $p(y = 1|x) = \frac{g_1(x)}{g_0(x)+g_1(x)}$ and $p(y = 0|x) = \frac{g_0(x)}{g_0(x)+g_1(x)}$ with $g_1(x) = \sum_i \frac{y_i}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)$, $g_0(x) = \sum_i \frac{1-y_i}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)$, and $\sigma = 0.5$. (b) The kernels $\mathbf{C}(x, 1, x_i, 1)$ (thin lines), the function $g_1(x) = \sum_i \mathbf{C}(x, 1, x_i, 1)$ (dashed line), and the likelihood $p(y = 1|x)$ (thick line).

The optimal reconstruction variance σ^2 is easily determined by cross–validation techniques. To reduce the numerical costs, the summation over i in Eq. (C.95) can be restricted to data points in the neighborhood of x or to a smaller number of terms at optimally adapted locations. The reconstruction variance σ^2 can be adapted locally, similar to nearest neighbor methods, for example, by taking into account the local density of the x–data.

Bibliography

[1] E. Aarts and J. Korts. *Simulated Annealing and Boltzmann Machines.* Wiley, New York, 1989.

[2] Y. Abu-Mostafa. Learning from hints in neural networks. *Journal of Complexity*, 6:192–198, 1990.

[3] Y. Abu-Mostafa. Hints and the VC dimension. *Neural Computation*, 5:278–288, 1993.

[4] Y. Abu-Mostafa. A method for learning from hints. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 73–80. Morgan Kaufmann, San Mateo, CA, 1993.

[5] A. Agresti. *Categorical Data Analysis.* Wiley, New York, 2nd edition, 2002.

[6] T. Aida. Field theoretical analysis of on-line learning of probability distributions. *Phys. Rev. Lett.*, 83:3554–3557, 1999. arXiv:cond-mat/9911474.

[7] R. G. Airapetyan and I. V. Puzynin. Newtonian iterative scheme with simultaneous iterations of inverse derivative. *Comp. Phys. Comm.*, 102:97–108, 1997.

[8] H. Akaike. Likelihood and the Bayes procedure. In J. M. Bernanda, M. H. De Groot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 143–166. Valencia University Press, Valencia, 1980.

[9] I. Aleksander, editor. *Neural Computing Architectures.* North Oxford Academic, London, 1989.

[10] I. Aleksander and H. Morton. *An Introduction to Neural Computing.* Chapman & Hall, London, 1990.

[11] I. Aleksander and H. Morton. *Neurons and Symbols.* Chapman & Hall, London, 1993.

[12] D. M. Allen. The relationship between variable selection and data augmentation and a method of prediction. *Technometrics*, 16:125, 1974.

[13] S.-I. Amari. *Differential Geometrical Methods in Statistics*, volume 28 of *Springer Lecture Notes in Statistics*. Springer–Verlag, New York, 1985.

[14] S.-I. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[15] S.-I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao, editors. *Differential Geometry in Statistical Inference*, volume 10 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1987.

[16] S.-I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, Cambridge, MA, 1996.

[17] W. F. Ames. *Numerical Methods for Partial Differential Equations*. Academic Press, New York, 2nd edition, 1977.

[18] T. G. Andersen and B. Sørensen. GMM estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Business and Economic Statistics*, 14:328–352, 1996.

[19] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 686:337–404, 1950.

[20] G. Auletta. *Foundations and Interpretations of Quantum Mechanics*. World Scientific, Singapore, 2000.

[21] A. Baddeley and R. Turner. Practical maximum pseudolikelihood for spatial point patterns. Research report 98-16, Department of Mathematics, University of Western Australia, August 1998. To appear in Australian and New Zealand Journal of Statistics as a discussion paper.

[22] V. Balasubramanian. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, 9(2):349–368, 1997.

[23] R. Balian. *From Microphysics to Macrophysics*, volume 1. Springer–Verlag, Berlin, 1991.

[24] D. H. Ballard. *An Introduction to Natural Computation*. MIT Press, Cambridge, MA, 1997.

[25] K. Banaszek, G. M. D'Ariano, M. G. A. Paris, and M. F. Sacchi. Maximum-likelihood estimation of the density matrix. *Phys. Rev. A*, 61:R10304, 1999.

[26] O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* Wiley, New York, 1978.

[27] T. R. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society London*, 53:370, 1763. Reprinted in *Biometrika* 45:293, 1958.

[28] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming.* Wiley, New York, 2nd edition, 1993.

[29] C. Beck and F. Schlögl. *Thermodynamics of Chaotic Systems.* Cambridge University Press, Cambridge, 1993.

[30] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[31] J. S. Bell. *Speakable and Unspeakable in Quantum Mechanics.* Cambridge University Press, Cambridge, 1987.

[32] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications.* Wiley, New York, 1974.

[33] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer–Verlag, New York, 2nd edition, 1985.

[34] J. O. Berger and L. M. Berliner. Robust Bayes and empirical Bayes analysis with ϵ–contaminated priors. *Ann. Statist.*, 14:461–486, 1986.

[35] J. O. Berger and R. Wolpert. *The Likelihood Principle*, volume 9 of *Monograph Series.* IMS Lecture Notes, Hayward, CA, 2nd edition, 1988.

[36] P. L. Berger and T. Luckmann. *Die gesellschaftliche Konstruktion der Wirklichkeit.* S. Fischer Verlag, Frankfurt/Main, 1969.

[37] J. M. Bernado. Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society A*, 41:113–147, 1979.

[38] J. M. Bernado and A. F. Smith. *Bayesian Theory.* Wiley, New York, 1994.

[39] M. Bertero, T. Poggio, and V. Torre. Ill–posed problems in early vision. In *Proceedings of the IEEE*, volume 76, pages 869–889, 1988.

[40] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1995.

[41] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36:192–236, 1974.

[42] J. Besag. Statistical analysis of non–lattice data. *The Statistician*, 24:179–195, 1975.

[43] J. Besag. Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute*, 47:77–92, 1977.

[44] D. Beymer, A. Shashua, and T. Poggio. Example based image analysis and synthesis. Artificial Intelligence Laboratory Memo 1431, Massachusetts Institute of Technology, Cambridge, MA, 1993.

[45] W. Bialek, C. G. Callan, and S. P. Strong. Field theories for learning probability distributions. *Phys. Rev. Lett.*, 77:4693–4697, 1996. arXiv:cond-mat/9607180.

[46] B. M. Bibby and M. Sørensen. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1:17–39, 1995.

[47] K. Binder, editor. *Monte Carlo Method in Condensed Matter Physics.* Springer–Verlag, Berlin, 1988.

[48] K. Binder and D. W. Heermann. *Monte Carlo Simulation in Statistical Physics: An Introduction.* Springer–Verlag, Berlin, 2nd edition, 1988.

[49] K. Binder and D. W. Heermann. *Monte Carlo Simulation in Statistical Physics.* Springer–Verlag, Berlin, corr. 2nd edition, 1992.

[50] S. J. Biondo. *Fundamentals of Expert Systems Technology: Principles and Concepts.* Ablex, Norwood, NJ, 1990.

[51] C. M. Bishop. Curvature–driven smoothing: A learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, 4(5):882–884, 1993.

[52] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[53] C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

[54] E. Bishop and D. Bridges. *Constructive Analysis*, volume 279 of *Grundlehren der Mathematischen Wissenschaften.* Springer–Verlag, Berlin, 1985.

[55] M. J. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Int'l J. Computer Vision*, 19(1), 1996.

[56] J.-P. Blaizot and G. Ripka. *Quantum Theory of Finite Systems.* MIT Press, Cambridge, MA, 1986.

[57] A. Blake and A. Zisserman. *Visual Reconstruction.* MIT Press, Cambridge, MA, 1987.

[58] P. Blanchard and E. Bruening. *Variational Methods in Mathematical Physics.* Springer–Verlag, Berlin, 1982.

[59] N. Bleistein and N. Handelsman. *Asymptotic Expansions of Integrals.* Dover, New York, 1986. Originally published in 1975 by Holt, Rinehart, and Winston, New York.

[60] P. Boehner, editor. *Ockham: Philosophical Writings.* Nelson, Edinburgh, 1957.

[61] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison–Wesley, Reading, MA, 1973. Wiley Classics Library Edition published 1992.

[62] L. Breiman. *Probability.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 1992. Originally published by Addison–Wesley, Reading, MA, 1968.

[63] L. Breiman. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. Inform. Theory*, 39(3):999–1013, 1993.

[64] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Chapman & Hall, New York, 1993.

[65] G. L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*, volume 48 of *Lecture Notes in Statistics.* Springer–Verlag, Berlin, 1988. Available at `http://bayes.wustl.edu/glb/book.pdf`.

[66] R. Brunelli and T. Poggio. Template matching: Matched spatial filters and beyond. Artificial Intelligence Laboratory Memo 1549, Massachusetts Institute of Technology, Cambridge, MA, 1995.

[67] B. Buck and V. A. Macaulay, editors. *Maximum Entropy in Action.* Oxford University Press, Oxford, 1991.

[68] W. Buntine and A. Weigend. Bayesian back–propagation. *Complex Systems*, 5:603, 1991.

[69] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[70] K. Burrage, A. Willimas, J. Erhel, and B. Pohl. Implementation of a generalized cross validation algorithm using deflation techniques for systems of linear equations. *Appl. Num. Math.*, 19:17–31, 1995.

[71] V. Buzek, G. Drobný, R. Derka, G. Adam, and H. Wiedemann. Reconstruction of quantum states from incomplete data. *Chaos, Solitons, and Fractals*, 10:981–1074, 1999. arXiv:quant-ph/9805020.

[72] J. Cardy. *Scaling and Renormalization in Statistical Physics.* Cambridge University Press, Cambridge, 1996.

[73] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*, volume 69 of *Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC, Boca Raton, FL, 1996.

[74] N. Cartwright. *How the Laws of Physics Lie.* Clarendon Press, Oxford, 1983.

[75] N. Cartwright. *Nature's Capacities and Their Measurement.* Clarendon Press, Oxford, 1989.

[76] K. Chadan, D. Colton, L. Päivärinta, and W. Rundell. *An Introduction to Inverse Scattering and Inverse Spectral Problems.* SIAM, Philadelphia, 1997.

[77] K. Chadan and P. C. Sabatier. *Inverse Problems in Quantum Scattering Theory.* Springer–Verlag, New York, 1989.

[78] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34:305–334, 1987.

[79] C. B. Chiu, E. C. G. Sudarshan, and B. Misra. Time evolution of unstable quantum states and a resolution of Zeno's paradox. *Phys. Rev. D*, 16(2):520–529, 1976.

[80] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick. *Analysis, Manifolds, and Physics. Part I: Basics.* North–Holland, Amsterdam, 1982.

[81] A. Cichocki and R. Unbehauen. *Neural Networks for Optimization and Signal Processing.* Wiley, Chichester, 1993.

[82] B. S. Clarke and A. R. Barron. Information–theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[83] P. Clifford. Markov random fields in statistics. In G. R. Grimmet and D. J. A. Welsh, editors, *Disorder in Physical Systems. A Volume in Honour of John M. Hammersley*, pages 19–32. Clarendon Press, Oxford, 1990.

[84] J. Collins. *Renormalization.* Cambridge University Press, Cambridge, 1984.

[85] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

[86] D. R. Cox and D. V. Hinkley. *Theoretical Statistics.* Chapman & Hall, London, 1974.

[87] R. T. Cox. Probability, frequency, and reasonable expectation. *Am. J. Phys.*, 14:1, 1946.

[88] R. T. Cox. *The Algebra of Probable Inference.* Johns Hopkins University Press, Baltimore, 1961.

[89] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross–validation. *Numer. Math.*, 31:377–403, 1979.

[90] N. A. C. Cressie. *Statistics for Spatial Data.* Wiley, New York, 1993.

[91] M. Creutz. *Quarks, gluons, and lattices.* Cambridge University Press, Cambridge, 1983.

[92] N. Cristianini, C. Campbell, and J. Shawe-Taylor. Dynamically adapting kernels in support vector machines. In M. S. Kearns, S. S. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 204–210. MIT Press, Cambridge, MA, 1999.

[93] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel–Based Learning Methods.* Cambridge University Press, Cambridge, 2000.

[94] F. Cucker and S. Smale. On the mathematical foundation of learning. *Bulletin of the American Mathematical Society*, 39(1):1–39, 2002.

[95] G. D'Agostini. Bayesian reasoning in high energy physics: Principles and applications. Yellow Report 99–03, CERN, 1999. Available at `http://www-zeus.roma1.infn.it/~agostini/prob+stat.html`.

[96] A. P. David. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2:25–36, 1992.

[97] R. Davidson and J. G. MacKinnon. *Estimation and Inference in Econometrics.* Oxford University Press, New York, 1993.

[98] L. Davis, editor. *Genetic Algorithms and Simulated Annealing.* Morgan Kaufmann, San Mateo, CA, 1987.

[99] L. Davis, editor. *Handbook of Genetic Algorithms.* Van Nostrand Reinhold, New York, 1991.

[100] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The Helmholtz machine. *Neural Computation*, 7:889–904, 1995.

[101] N. G. de Bruijn. *Asymptotic Methods in Analysis.* Dover, New York, 1981. Originally published by North–Holland Publishing Co., Amsterdam, 1958.

[102] B. de Finetti. *Probability, Induction, and Statistics.* Wiley, New York, 1972.

[103] G. Deco and D. Obradovic. *An Information–Theoretic Approach to Neural Computing.* Springer–Verlag, New York, 1996.

[104] A. P. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.

[105] H.-P. Deutsch. *Derivate und Interne Modelle.* Schäffer–Poeschel, Stuttgart, 2nd edition, 2001.

[106] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer–Verlag, New York, 1996.

[107] C. Di Castro and G. Jona-Lasinio. The renormalization group approach to critical phenomena. In C. Domb and M. S. Green, editors, *Phase Transitions and Critical Phenomena*, volume 6. Academic Press, London, 1976.

[108] P. Diaconis and D. Freedman. Asymptotic of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984.

[109] R. Dietrich, M. Opper, and H. Sompolinsky. Statistical mechanics of support vector networks. *Phys. Rev. Lett.*, 82(14):2975–2978, 1999.

[110] D. L. Donoho and I. M. Johnstone. Projection–based approximation and a duality with kernel methods. *Ann. Statist.*, 17(1):58–106, 1989.

[111] J. L. Doob. *Stochastic Processes.* Wiley, New York, 1953. New edition published 1990.

[112] V. Dose and W. von der Linden. Outlier tolerant parameter estimation. In W. von der Linden, V. Dose, R. Fischer, and R. Preuss, editors, *Maximum Entropy and Bayesian Methods.* Kluwer Academic Publishers, Dordrecht, 1999.

[113] R. M. Dudley. *A Course on Empirical Processes*, volume 1097 of *Lecture Notes in Mathematics,* pages 2–142. Springer–Verlag, New York, 1984.

[114] J. Durbin and S. J. Koopman. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84(3):669–684, 1997.

[115] W. Ebeling, J. Freund, and F. Schweitzer. *Komplexe Strukturen: Entropie und Information.* Teubner, Stuttgart, 1998.

[116] E. Eells. *Probabilistic Causality.* Cambridge University Press, Cambridge, 1991.

[117] K. B. Efetov. Supersymmetry and theory of disordered metals. *Advances in Physics*, 32:53–127, 1983.

[118] B. Efron. Bootstrap methods: Another look at the jacknife. *Ann. Statist.*, 7:1–26, 1979.

[119] B. Efron. Nonparametric estimates of the standard error: The jackknife, the bootstrap, and other methods. *Biometrika*, 68:589–599, 1981.

[120] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, New York, 1993.

[121] J. M. Eisenberg and W. Greiner. *Microscopic Theory of the Nucleus.* North–Holland, Amsterdam, 1972.

[122] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer, Cordrecht, 1996.

[123] B. S. Everitt and D. J. Hand. *Finite Mixture Distributions.* Chapman & Hall, New York, 1981.

[124] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

[125] R. Fernández, J. Fröhlich, and A. D. Sokal. *Random Walks, Critical Phenomena, and Triviality in Quantum Field Theory.* Springer–Verlag, Berlin, 1992.

[126] H. Feshbach. *Theoretical Nuclear Physics: Nuclear Reactions.* Wiley, New York, 1992.

[127] M. W. Firebaugh. *Artificial Intelligence: A Knowledge–Based Approach.* PWS–Kent, Boston, MA, 1989.

[128] K. H. Fischer and J. A. Hertz. *Spin Glasses.* Cambridge University Press, Cambridge, 1991.

[129] M. C. Fischer, B. Gutiérrez-Medina, and M. G. Raizen. Observation of the quantum Zeno and Anti–Zeno effects in an unstable system. *Phys. Rev. Lett.*, 87(4):040402, 2001.

[130] R. Fischer, K. M. Hanson, V. Dose, and W. von der Linden. Background estimation in experimental spectra. *Phys. Rev. E*, 61:1152, 2000.

[131] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309, 1922.

[132] R. A. Fisher. Inverse probability. *Proc. Camb. Phil. Soc.*, 26:528, 1930.

[133] R. A. Fisher. *Statistical Methods and Scientific Inference.* Oliver and Boyd, London, 2nd edition, 1965.

[134] R. Fletcher. *Practical Methods of Optimization.* Wiley, New York, 1987.

[135] L. Fonda, G. C. Ghirardi, and A. Rimini. Decay theory of unstable quantum systems. *Rep. Prog. Phys.*, 41:587–631, 1978.

[136] I. Fredholm. Sur une classe d'équations fonctionelles. *Acta Math.*, 27:365–390, 1903.

[137] D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–316, 2001.

[138] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Am. Statist. Assoc.*, 76(376):817–823, 1981.

[139] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, 24:1000–1006, 1974.

[140] K. Fukunaga. *Statistical Pattern Recognition.* Academic Press, Boston, 1990.

[141] S. I. Gallant. *Neural Network Learning.* MIT Press, Cambridge, MA, 1993.

[142] C. W. Gardiner. *Handbook of Stochastic Methods.* Springer–Verlag, Berlin, 2nd edition, 1990.

[143] E. Gardner. Maximum storage capacity in neural networks. *Europhysics Letters*, 4:481–485, 1987.

[144] E. Gardner. The space of interactions in neural network models. *Journal of Physics A*, 21:257–270, 1988.

[145] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A*, 21:271–284, 1988.

[146] D. Geiger and F. Girosi. Parallel and deterministic algortihms for MRFs: Surface reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(5):401–412, 1991.

[147] D. Geiger and A. L. Yuille. A common framework for image segmentation. *Int'l J. Computer Vision*, 6(3):227–243, 1991.

[148] I. M. Gel'fand and B. M. Levitan. On the determination of a differential equation from its spectral function. *Trans. Amer. Soc.*, 1:253–302, 1951.

[149] S. B. Gelfand and S. K. Mitter. *On Sampling Methods and Annealing Algorithms. Markov Random Fields - Theory and Applications.* Academic Press, New York, 1993.

[150] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, New York, 1995.

[151] D. Geman and G. Reynolds. Constraint restoration and the recover of discontinuities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14:367–383, 1992.

[152] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984. Reprinted in Shafer & Pearl, editors, 1990. *Readings in Uncertainty Reasoning.* Morgan Kaufmann, San Mateo, CA, 1990.

[153] C. Geyer. Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, 7:473–511, 1992.

[154] M. Gibbs and D. J. C. MacKay. Efficient implementation of Gaussian processes. Cavendish Laboratory, Cambridge, 1997. Draft manuscript, available from `http://www.cs.toronto.edu/ mackay/gpros.ps.gz`.

[155] F. J. Giessibl, S. Hembacher, H. Bielefeld, and J. Mannhart. Subatomic features on the silicon (111)-(7×7) surface observed by atomic force microscopy. *Science*, 289:422–425, 2000.

[156] T. Gilovich, D. Griffin, and D. Kahnemann, editors. *Heuristics and Biases.* Cambridge University Press, Cambridge, 2002.

[157] B. G. Giraud, A. Lapedes, L. C. Liu, and J. C. Lemm. Lorentzian neural nets. *Neural Networks*, 8(5):757–767, 1995.

[158] F. Girosi. Models of noise and robust estimates. Artificial Intelligence Laboratory Memo 1287, Massachusetts Institute of Technology, Cambridge, MA, 1991.

[159] F. Girosi. Some extensions of radial basis functions and their applications in artificial intelligence. *Computers Math. Applic.*, 24(12):61–80, 1992.

[160] F. Girosi. An equivalence between sparse approximation and support vector machines. Artificial Intelligence Laboratory Memo 1606, Massachusetts Institute of Technology, Cambridge, MA, 1997.

[161] F. Girosi and N. T. Chan. Prior knowledge and the creation of virtual examples for rbf networks. In *IEEE Workshop on Neural Networks for Signal Processing*, Cambridge, MA, 1995.

[162] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural network architectures. *Neural Computation*, 7(2):219–269, 1995.

[163] F. Girosi, T. Poggio, and B. Caprile. Extensions of a theory of networks for approximations and learning: Outliers and negative examples. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems 3*. Morgan Kaufmann, San Mateo, CA, 1991.

[164] D. Giulini, E. Joos, C. Kiefer, I.-O. Stamatescu, and H. D. Zeh. *Decoherence and the Appearance of a Classical World in Quantum Theory*. Springer–Verlag, Berlin, 1996.

[165] J. Glimm and A. Jaffe. *Quantum Physics. A Functional Integral Point of View*. Springer–Verlag, New York, 1987.

[166] K. Goeke, R. Y. Cusson, F. Grümmer, P.-G. Reinhard, and H. Reinhardt. Time–dependent Hartree–Fock and beyond: A review. *Prog. Theor. Phys. (Suppl.)*, 74 & 75:33, 1983.

[167] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison–Wesley, Redwood City, CA, 1989.

[168] R. M. Golden. *Mathematical Methods for Neural Network Analysis and Design*. MIT Press, Cambridge, MA, 1996.

[169] G. H. Golub, M. Heath, and G. Wahba. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–224, 1979.

[170] G. H. Golub and C. F. van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, 3rd edition, 1996.

[171] G. H. Golub and U. von Matt. Generalized cross–validation for large scale problems. *Journal of Computational and Graphical Statistics,* 6(1):1–34, 1997.

[172] J. Gomes and L. Velho. *Image Processing for Computer Graphics.* Springer–Verlag, New York, 1997.

[173] R. C. Gonzales and R. E. Woods. *Digital Image Processing.* Addison–Wesley, Reading, MA, 1993.

[174] I. J. Good. Some history of the hierachical Bayesian methodology. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics II.* North–Holland, Amsterdam, 1980.

[175] I. J. Good. *Good Thinking: The Foundations of Probability and Its Applications.* University of Minnesota Press, Minneapolis, 1983.

[176] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika,* 58:255–277, 1971.

[177] C. Goodall. M–estimators of location: An outline of the theory. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors, *Understanding Robust and Exploratory Data Analysis,* pages 339–403. Wiley, New York, 1983.

[178] G. Görz, editor. *Einführung in die künstliche Intelligenz.* Addison–Wesley, Bonn, Paris, 1993.

[179] C. Gouriéroux. *ARCH Models and Financial Applications.* Springer–Verlag, New York, 1997.

[180] C. Gouriéroux, A. Monfort, and A. Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica,* 52:681–700, 1984.

[181] W. P. Gouveia and J. A. Scales. Resolution of seismic waveform inversion: Bayes versus Occam. *Inverse Problems,* 13:323–349, 1997.

[182] I. Gradstein and I. Ryshik. *Tables of Series, Products, and Integrals.* Harry Deutsch, 1981.

[183] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models,* volume 58 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London, 1994.

[184] C. Großman and H.-G. Roos. *Numerik partieller Differentialgleichungen.* Teubner, Stuttgart, 1994.

[185] S. F. Gull. Bayesian data analysis – straight line fitting. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 511–518. Kluwer Academic Publishers, Dordrecht, 1989.

[186] S. F. Gull. Developments in maximum entropy data analysis. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 53–71. Kluwer Academic Publishers, Dordrecht, 1989.

[187] S. F. Gull and G. J. Daniell. The maximum entropy method. In C. Van Schooneveld, editor, *Image formation from coherence functions in astronomy*, pages 219–225. Reidel, Dordrecht, 1979.

[188] H. Gumin and Meier H., editors. *Einführung in den Konstruktivismus.* Piper, München, Zürich, 1992.

[189] W. Hackbusch. *Multi-grid Methods and Applications.* Springer–Verlag, New York, 1985.

[190] W. Hackbusch. *Integralgleichungen.* Teubner, Stuttgart, 1989.

[191] W. Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* Teubner, Stuttgart, 1993.

[192] J. Hadamard. *Lectures on the Cauchy Problem in Linear Partial Differential Equations.* Yale University Press, New Haven, 1923.

[193] P. Hall. *The Bootstrap and Edgeworth Expansion.* Springer–Verlag, New York, 1992.

[194] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished, 1971.

[195] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*, volume 3 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London, 1964.

[196] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics. The Approach Based on Influence Functions.* Wiley, New York, 1986.

[197] L. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

[198] W. Härdle. *Applied nonparametric regression.* Cambridge University Press, Cambridge, 1990.

[199] J. Hartung. *Statistik.* Oldenbourg, München, 13th edition, 2002.

[200] A. C. Harvey, E. Ruiz, and N. Shepard. Multivariate stochastic volatility models. *Review of Econometric Studies*, 61:247–264, 1994.

[201] T. J. Hastie and R. J. Tibshirani. Generalized additive models. *Statist. Sci.*, 1:297–318, 1986.

[202] T. J. Hastie and R. J. Tibshirani. Generalized additive models: Some applications. *J. Am. Statist. Assoc.*, 82:371–386, 1987.

[203] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1990.

[204] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[205] F. Hayes-Roth. The knowledge based expert system: A tutorial. *IEEE Computer*, 17(9):11–28, 1984.

[206] R. Hecht-Nielsen. *Neurocomputing*. Addison–Wesley, Reading, MA, 1990.

[207] C. W. Helstrom. *Quantum Detection and Estimation Theory*. Academic Press, New York, 1976.

[208] R. Herbrich, T. Graepel, and C. Campell. Bayesian learning in reproducing kernel Hilbert spaces. Technical Report TR 99–11, Technische Universität Berlin, 1999. Available from `http://stat.cs.tu-berlin.de/~ralfh/publications.html`.

[209] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computation*. Addison–Wesley, Redwood City, CA, 1991.

[210] D. Hilbert and R. Courant. *Methods of Mathematical Physics*, volumes 1 & 2. Wiley, New York, 1989. First German editions published by Springer–Verlag, Berlin, 1924, 1937.

[211] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The 'wake–sleep' algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995.

[212] G. E. Hinton and T. J. Sejnowski. Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing*, pages 282–317. MIT Press, Cambridge, MA, 1986.

[213] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors. *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, 1983.

[214] H. Hochstadt and B. Lieberman. An inverse Sturm–Liouville problem with mixed given data. *SIAM J. Appl. Math.*, 34:676–680, 1976.

[215] R. Hoffmann, V. I. Minkin, and B. K. Carpenter. Ockham's razor and chemistry. *Bulletin de la Société Chimique de France*, 133:117–130, 1996. Available at `http://www.rz.uni-karlsruhe.de/ ed01/Hyle/Hyle3/hoffman.htm`.

[216] B. Hofmann. *Mathematik inverser Probleme.* Teubner Verlag, Stuttgart, Leibzig, 1999.

[217] A. S. Holevo. *Probabilistic and Statistical Aspects of Quantum Theory.* North–Holland, Amsterdam, 1982.

[218] J. H. Holland. *Adaption in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, MC, 1975. MIT Press, Cambridge, MA, second edition, 1992.

[219] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[220] J. Honerkamp. *Stochastic Dynamical Systems. Concepts, Numerical Methods, Data Analysis.* VCH Verlag, New York, 1994.

[221] J. Honerkamp. *Statistical Physics.* Springer–Verlag, New York, 1998.

[222] J. Honerkamp and J. Weese. Tikhonovs regularization method for ill–posed problems: A comparison of different methods. *Cont. Mech. Thermodyn.*, 2:17–30, 1990.

[223] R. Horst, M. Pardalos, and N. V. Thoai. *Introduction to Global Optimization.* Kluwer Academic Publishers, Dordrecht, 1995.

[224] P. J. Huber. Robust estimation of location parameter. *Ann. Math. Stat.*, 35(1):1753–1758, 1964.

[225] P. J. Huber. Robust smoothing. In E. Launer and G. Wilkinson, editors, *Robustness in Statistics.* Academic Press, New York, 1979.

[226] P. J. Huber. *Robust Statistics.* Wiley, New York, 1981.

[227] P. J. Huber. Projection pursuit. *Ann. Statist.*, 13(2):435–475, 1985.

[228] J. Hull. *Options, Futures, and Other Derivative Securities.* Prentice Hall, Englewood Cliffs, NJ, 1989.

[229] B. Iglewicz. Robust scale estimators and confidence intervals for location. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors, *Understanding Robust and Exploratory Data Analysis*, pages 405–431. Wiley, New York, 1983.

[230] J. P. Ignizio. *Introduction to Expert Systems: The Development and Implementation of Rule-Based Expert Systems.* Prentice Hall, Englewood Cliffs, NJ, 1989.

[231] W. M. Itano, D. J. Heinzen, J. J. Bollinger, and D. J. Wineland. Quantum Zeno effect. *Phys. Rev. A*, 41:2295–2300, 1990.

[232] C. Itzkyson and J.-M. Drouffe. *Statistical Field Theory,* volumes 1 & 2. Cambridge University Press, Cambridge, 1989.

[233] B. Jähne. *Digitale Bildverarbeitung.* Springer–Verlag, Berlin, 4th edition, 1997.

[234] E. T. Jaynes. Probability theory: The logic of science, 2000. Available at `http://bayes.wustl.edu/etj/prob.html`.

[235] R. C. Jeffrey. *The Logic of Decision.* University of Chicago Press, Chicago, 1983.

[236] R. C. Jeffrey. Probabilistic judgment, 1999. Available from `http://www.princeton.edu/~bayesway/`.

[237] R. C. Jeffrey. Probabilistic thinking, 1999. Available from `http://www.princeton.edu/~bayesway/`.

[238] H. Jeffreys. *Theory of Probability.* Oxford University Press, Oxford, 3rd edition, 1961. First edition published 1939.

[239] H. Jeffreys and J. O. Berger. Ockham's razor and Bayesian analysis. *American Scientist*, 80:468–477, 1992.

[240] H. Jeggle. *Nichtlineare Funktionalanalys.* Teubner, Stuttgart, 1979.

[241] F. V. Jensen. *An Introduction to Bayesian Networks.* Springer–Verlag, New York, 1996.

[242] F. V. Jensen, S. L. Lauritzen, and K. G. Olesen. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly*, 5:269–282, 1990.

[243] M. C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society A*, 150:1–36, 1987.

[244] M. J. Jones and T. Poggio. Model-based matching of line drawings by linear combinations of prototypes. In *Proceedings of the IEEE 5th International Conference on Computer Vision*, pages 531–536. IEEE Computer Society Press, Cambridge, MA, 1995.

[245] M. Jordan, editor. *Learning in Graphical Models.* MIT Press, Cambridge, MA, 1999.

[246] M. I. Jordan and R. A. Jacobs. Hierachical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[247] M. Kac. Can one hear the shape of a drum? *Am. Math. Mon.*, 73:1–23, 1966.

[248] D. Kahneman, P. Slovic, and A. Tversky. *Judgement Under Uncertainty: Heuristics and Biases.* Cambridge University Press, New York, 1982.

[249] L. Kahnemann and A. Tversky. Judgment under uncertainty: Heuristics and biases. *Science*, 182:1124–1131, 1974.

[250] L. Kahnemann and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(263):269–291, 1979.

[251] M. Kaku. *Quantum Field Theory.* Oxford University Press, Oxford, 1993.

[252] A. Kandel. *Fuzzy Techniques in Pattern Recognition.* Wiley, New York, 1982.

[253] I. Kant. *Kritik der reinen Vernunft*, volume 3 of *Kants Werke.* Königlich Preußische Akademie der Wissenschaften, Berlin, 2nd edition, 1787. Reprinted by Walter de Gruyter & Co, Berlin, 1968.

[254] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, Cambridge, MA, 1994.

[255] J. B. Keller. Inverse problems. *Ann. Math. Mon.*, 83:107–118, 1976.

[256] D. F. Kerridge. Inaccuracy and inference. *Journal of the Royal Statistical Society*, 23:184–194, 1961.

[257] J. M. Keynes. *A Treatise on Probability.* MacMillan, London, 1921.

[258] S. Kim, N. Shephard, and S. Chib. Stochastic volatility: Likelihood inference and comparison with arch models. *Review of Economic Studies*, 65:361–393, 1998.

[259] G. S. Kimmeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41:495–502, 1970.

[260] G. S. Kimmeldorf and G. Wahba. Spline functions and stochastic processes. *Sankhya Ser. A*, 32(Part 2):173–180, 1970.

[261] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[262] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems.* Springer–Verlag, New York, 1996.

[263] G. Kitagawa and W. Gersch. *Smoothness Priors Analysis of Time Series.* Springer–Verlag, New York, 1996.

[264] H. Kleinert. *Pfadintegrale.* Wissenschaftsverlag, Mannheim, 1993.

[265] G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic.* Prentice Hall, Upper Saddle River, NJ, 1995.

[266] G. J. Klir and B. Yuan, editors. *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems.* World Scientific, Singapore, 1996.

[267] M. Koecher. *Lineare Algebra und analytische Geometrie.* Springer–Verlag, Berlin, 1985.

[268] A. G. Kofman and G. Kurizki. Acceleration of quantum decay processes by frequent observations. *Nature*, 405:546–549, 2000.

[269] T. Kohonen. *Self–Organizing and Associative Memory.* Springer, Berlin, 1984.

[270] T. Kohonen. *Self–Organizing Maps.* Springer, Berlin, 1995.

[271] P. Kontkanen, P. Myllymäki, and H. Tirri. Constructing Bayesian finite mixture models by the EM algorithm. Technical Report Series NC–TR–97–003, NeuroCOLT, 1997.

[272] H. Kopp. *Bildverarbeitung interaktiv.* Teubner, Stuttgart, 1997.

[273] B. Kosko. *Neural Networks for Signal Processing.* Prentice–Hall, Englewood Cliffs, NJ, 1992.

[274] J. R. Koza. *Genetic Programming.* MIT Press, Cambridge, MA, 1992.

[275] S. Kullback. *Information Theory and Statistics.* Wiley, New York, 1951.

[276] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[277] H. Kyburg Jr. and H. Smokler, editors. *Studies in Subjective Probabilities.* Robert E. Krieger, Huntington, NY, 1980.

[278] D. P. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics.* Cambridge University Press, Cambridge, 2000.

[279] A. Lapedes and R. Farber. How neural nets work. In D. Z. Anderson, editor, *Neural Information Processing Systems*, pages 442–456. American Institute of Physics, New York, 1988.

[280] P. S. Laplace. *Theorie Analytique des Probabilités.* Courcier, Paris, 1812.

[281] P. S. Laplace. *Philosophical Essay on Probability.* Dover, New York, 1951.

[282] S. L. Lauritzen. *Graphical Models.* Oxford University Press, Oxford, 1996.

[283] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems.* Prentice Hall, Englewood Cliffs, NJ, 1974.

[284] M. Le Bellac. *Quantum and Statistical Field Theory.* Oxford Science Publications, Clarendon Press, Oxford, 1991.

[285] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer–Verlag, New York, 1986.

[286] T. K. Leen. From data distributions to regularization in invariant learning. *Neural Computation*, 7(5):974–981, 1995.

[287] J. C. Lemm. Lecture notes 'Econophysics'. Available at `http://pauli.uni-muenster.de/~lemm/econoWS99/econo.ps.gz`.

[288] J. C. Lemm. Entscheidungsstrategien und Informationskosten. Diplomarbeit, Fachbereich Psychologie, Universität Münster, 1984.

[289] J. C. Lemm. Inhomogeneous random phase approximation: A solvable model. *Annals of Physics*, 244(1):201–238, 1995.

[290] J. C. Lemm. Inhomogeneous random phase approximation for nuclear and atomic reactions. *Annals of Physics*, 244(1):136–200, 1995.

[291] J. C. Lemm. Prior information and generalized questions. Artificial Intelligence Laboratory Memo 1598, Massachusetts Institute of Technology, Cambridge, MA, 1996.

[292] J. C. Lemm. Fuzzy interface with prior concepts and non–convex regularization. In Wilfried Brauer, editor, *Fuzzy–Neuro Systems '98– Computational Intelligence. Proceedings of the 5. International Workshop 'Fuzzy-Neuro Systems '98' (FNS '98), March 19–20, 1998, Munich, Germany*, Proceedings in Artificial Intelligence, pages 298–305. Infix, Sankt Augustin, 1998.

[293] J. C. Lemm. Fuzzy rules and regularization theory. In ELITE European Laboratory for Intelligent Techniques Engineering, editor, *6th European Congress on Intelligent Techniques and Soft Computing (EUFIT '98).* Mainz, Aachen, 1998.

[294] J. C. Lemm. How to implement a priori information: A statistical mechanics approach. Universität Münster, Institut für Theoretische Physik, MS-TP1-98-12, 1998. arXiv:cond-mat/9808039.

[295] J. C. Lemm. Quadratic concepts. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *ICANN 98. Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, September 1–4, 1998*, pages 579–584. Springer–Verlag, London, 1998.

[296] J. C. Lemm. Bayesian field theory. Technical Report MS-TP1-99-1, Universität Münster, 1999. arXiv:physics/9912005.

[297] J. C. Lemm. Mixtures of Gaussian process priors. In *ICANN 99. Proceedings of the 9th International Conference on Artificial Neural Networks, Edinburgh, UK, September 7–10, 1999*. IEEE Conference Publication, London, 1999.

[298] J. C. Lemm. Inverse time–dependent quantum mechanics. *Phys. Lett. A*, 276(1–4):19–24, 2000. arXiv:quant-ph/0002010.

[299] J. C. Lemm. Nonparametric approaches to density estimation. In *Proceedings of the International Joint Conference on Neural Networks, Como, Italy*. IEEE Conference Publications, London, 2000.

[300] J. C. Lemm. Bayesian field theory and approximate symmetries. In A. Mohammad-Djafari, editor, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, International Workshop, Gif sur Yvette, France, 2000*, volume 568 of *AIP Conference Proceedings*. American Institute of Physics, 2001. MS-TP1-00-10.

[301] J. C. Lemm, V. Beiu, and J. G. Taylor. Density estimation as a preprocessing step for constructive algorithms. In Kappen B. and S. Gielen, editors, *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd SNN Neural Network Symposium. The Netherlands, Nijmegen, September 14–15, 1995*, pages 213–216. Springer–Verlag, Berlin, 1995.

[302] J. C. Lemm, B. G. Giraud, and A. Weiguny. Beyond the time–independent mean field theory for nuclear and atomic reactions: Inclusion of particle–hole correlations in a generalized random phase approximation. *Phys. Rev. Lett.*, 73:420–423, 1994.

[303] J. C. Lemm and J. Uhlig. Bayesian inverse quantum theory. *Few–Body Systems*, 29:25–52, 2000. arXiv:quant-ph/0006027.

[304] J. C. Lemm and J. Uhlig. Hartree–Fock approximation for inverse many-body problems. *Phys. Rev. Lett.*, 84:4517–4520, 2000. arXiv:nucl-th/9908056.

[305] J. C. Lemm, J. Uhlig, and A. Weiguny. A Bayesian approach to inverse quantum statistics. *Phys. Rev. Lett.*, 84:2068–2071, 2000. arXiv:cond-mat/9907013.

[306] J. C. Lemm, J. Uhlig, and A. Weiguny. Bayesian reconstruction of approximately periodic potentials for quantum systems at finite temperatures. *The European Physical Journal B*, 20(3):349–366, 2001. MS-TP1-00-4, arXiv:quant-ph/0005122.

[307] H. Lenk. *Interpretation und Realität.* Suhrkamp, Frankfurt a. M., 1995.

[308] T. Leonard and J. S. J. Hsu. *Bayesian Methods.* Cambridge University Press, Cambridge, 1999.

[309] M. A. Lifshits. *Gaussian Random Functions.* Kluwer Academic Publishers, Dordrecht, 1995.

[310] D. V. Lindley. A statistical paradox. *Biometrica*, 44:187–192, 1957.

[311] D. V. Lindley. *Bayesian Statistics, A Review.* SIAM, Philadelphia, 1972.

[312] D. V. Lindley and L. D. Phillips. Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.*, 30:112–119, 1976.

[313] N. K. Logothetis, T. Vetter, A. Hurlbert, and T. Poggio. View–based models of 3d object recognition and class–specific invariances. Artificial Intelligence Laboratory Memo 1472, Massachusetts Institute of Technology, Cambridge, MA, 1994.

[314] T. Loredo. From Laplace to Supernova SN 1987A: Bayesian inference in astrophysics. In P. F. Fougère, editor, *Maximum-Entropy and Bayesian Methods, Dartmouth, 1989*, pages 81–142. Kluwer Academic Publishers, Dordrecht, 1990.
Available from `http://bayes.wustl.edu/gregory/gregory.html`.

[315] A. K. Louis. *Inverse und schlecht gestellte Probleme.* Teubner, Stuttgart, 1989.

[316] D. R. Luce and H. Raiffa. *Games and Decisions.* Wiley, New York, 1957. Also published by Dover, New York, 1989.

[317] D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

[318] D. J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[319] D. J. C. MacKay. Bayesian methods for backpropagation networks. In E. Domany, J. L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks III*, pages 211–254. Springer–Verlag, New York, 1996.

[320] D. J. C. MacKay. Bayesian non–linear modeling for the prediction competition. In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, pages 221–234. Kluwer Academic Publishers, Dordrecht, 1996.

[321] D. J. C. MacKay. Hyperparameters: Optimise or integrate out? In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, pages 43–59. Kluwer Academic Publishers, Dordrecht, 1996. An extended version appeared in *Neural Comp.* 11:1035-1068, 1999 under the title 'Comparison of Approximate Methods for Handling Hyperparameters'.

[322] D. J. C. MacKay. Introduction to Gaussian processes. In C. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *Series F, Computer and Systems Sciences*. NATO ASI Series, 1998.

[323] D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11:1035–1068, 1999. A preprint version is available at `http://wol.ra.phy.cam.ac.uk/mackay/alpha.ps.gz`.

[324] G. J. MacLachlan and K. E. Bashford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

[325] G. Mahler and V. A. Weberruß. *Quantum Networks*. Springer–Verlag, Berlin, 1995.

[326] C. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[327] B. B. Mandelbrot. *The Fractal Geometry of Nature*. Freeman and Company, New York, 1977.

[328] B. B. Mandelbrot. *Fractals and Scaling in Finance*. Springer–Verlag, New York, 1997.

[329] V. A. Marchenko. *Sturm–Liouville Operators and Applications*. Birkhauser, Basel, 1986.

[330] D. W. Marquardt. Generalized inverses, ridge regression, biased linear estimation, and nonlinear regression. *Technometrics*, 12:591–613, 1970.

[331] D. W. Marquardt and R. D. Snee. Ridge regression in practice. *The American Statistician*, 29:3–20, 1975.

[332] J. L. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill–posed problems in computational vision. *J. Am. Stat. Assoc.*, 82:76–89, 1987.

[333] P. McCullagh and J. A. Nelder. *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1989.

[334] M. L. Mehta. *Random Matrices*. Academic Press, San Diego, 2nd edition, 1991.

[335] M. Meila and M. I. Jordan. Learning with mixtures of trees. *J. of Machine Learning Research*, 1:1–48, 2000.

[336] A. Messiah. *Quantum Mechanics*. North–Holland, Amsterdam, 1961.

[337] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[338] M. Mezard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.

[339] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer–Verlag, Berlin, 1992.

[340] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York, 1994.

[341] D. Miller. *Critical Rationalism: A Restatement and Defence*. Open Court Publishing Company, Chicago, IL, 1994.

[342] R. G. Miller. *Simultaneous Statistical Inference*. McGraw–Hill, New York, 1966.

[343] W. T. Miller III, R. S. Sutton, and P. J. Werbos, editors. *Neural Networks for Control*. MIT Press, Cambridge, MA, 1990.

[344] M. L. Minski and S. A. Papert. *Perceptrons*. MIT Press, Cambridge, MA, expanded edition, 1990. Original edition published 1969.

[345] A. R. Mitchell and D. F. Griffiths. *The Finite Difference Method in Partial Differential Equations*. Wiley, New York, 1980.

[346] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, 1996.

[347] A. Mohammad-Djafari. A full Bayesian approach for inverse problems. In K. M. Hanson and R. N. Silver, editors, *Maximum Entropy and Bayesian Methods, Santa Fe, New Mexico, 1995.* Kluwer Academic Publishers, Dordrecht, 1996.

[348] A. Mohammad-Djafari and J. Idier. A scale invariant Bayesian method to solve linear inverse problems. In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, pages 121–134. Kluwer Academic Publishers, Dordrecht, 1996. An extended version can be found in the Santa Fe Institute Technical Report 94-07-001.

[349] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72(23):3634–3637, 1994.

[350] J. F. Monahan. *Numerical Methods of Statistics.* Cambridge University Press, Cambridge, 2001.

[351] I. Montvay and G. Münster. *Quantum Fields on a Lattice.* Cambridge University Press, Cambridge, 1994.

[352] E. H. Moore. On the reciprocal of the general algebraic matrix. *Bull. Amer. Math. Soc.*, 26:394–395, 1920.

[353] V. A. Morozov. *Methods for Solving Incorrectly Posed Problems.* Springer–Verlag, New York, 1984.

[354] C. Morris. Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.*, 78:47–65, 1983.

[355] P. M. Morse and H. Feshbach. *Methods of Theoretical Physics. Part II.* McGraw–Hill, New York, 1953.

[356] F. Mosteller and D. Wallace. Inference in an authorship problem. A comparative study of discrimination methods applied to authorships of the disputed Federalist papers. *J. Amer. Statist. Assoc.*, 58:275–309, 1963.

[357] C. Mudry, C. Chamon, and X.-G-Wen. Two–dimensional conformal field theory for disordered systems at criticality. *Nucl. Phys. B*, 466 [FS]:383–443, 1996.

[358] B. Müller and J. Reinhardt. *Neural Networks.* Springer–Verlag, Berlin, second printing, 1991.

[359] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Applied Math.*, 42:577–684, 1989.

[360] M. K. Murray and J. W. Rice. *Differential Geometry and Statistics*, volume 48 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1993.

[361] E. A. Nadaraya. On nonparametric estimates of density functions and regression curves. *Theor. Prob. Appl.*, 10:186–190, 1965.

[362] F. Natterer. *The Mathematics of Computerized Tomography*. Teubner and Wiley, Stuttgart, 1986.

[363] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer–Verlag, New York, 1996.

[364] R. M. Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report 9702, Department of Statistics, University of Toronto, Canada, 1997.

[365] J. W. Negele and H. Orland. *Quantum Many–Particle Systems*, volume 68 of *Frontiers in Physics Series*. Addison–Wesley, Redwood City, CA, 1988.

[366] R. G. Newton. *Scattering Theory of Waves and Particles*. Springer–Verlag, Berlin, second edition, 1982.

[367] R. G. Newton. *Inverse Schrödinger Scattering in Three Dimensions*. Springer–Verlag, Berlin, 1989.

[368] M. Nitzberg and T. Shiota. Nonlinear image filtering with edge and corner enhancement. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(8):862–833, 1992.

[369] P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virtual examples. In *IEEE Proceedings on Intelligent Signal Processing*, volume 86, pages 2196–2209, 1998.

[370] J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer–Verlag, New York, 1996.

[371] A. O'Hagen. *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Edward Arnold, London, 1994.

[372] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7:333–339, 1995. Proceedings of the Workshop on Information Theory and the Brain, September 4–5, 1995, University of Stirling.

[373] B. A. Olshausen and D. J. Field. Emergence of simple–cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[374] M. Opper. Gaussian processes for classification: Mean field algorithms. Technical Report NCRG/1999/030, Neural Computing Research Group at Aston University, 1999.

[375] M. Opper and D. Haussler. Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Phys. Rev. Lett.*, 66:2677, 1991.

[376] M. Opper and W. Kinzel. Statistical mechanics of generalization. In E. Domany, J. L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks III*, pages 151–210. Springer–Verlag, New York, 1996.

[377] M. Opper and O. Winther. Mean field methods for classification with Gaussian processes. In M.S. Kearns, S. S. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 309–315. MIT Press, Cambridge, MA, 1999.

[378] M. Opper and O. Winther. Gaussian process classification and SVM: Mean field results and leave–one–out estimator. In A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 43–65. MIT Press, Cambridge, MA, 2000.

[379] Y. Pao. *Adaptive Pattern Recognition and Neural Networks.* Addison–Wesley, Reading, MA, 1989.

[380] G. Parisi. *Field Theory, Disorder, and Simulations.* World Scientific, Singapore, 1992.

[381] E. Parzen. An approach to time series analysis. *Ann. Math. Statist.*, 32:951–989, 1962.

[382] E. Parzen. On the estimation of a probability function and mode. *Ann. Math. Statist.*, 33:3, 1962.

[383] E. Parzen. Probability density functionals and reproducing kernel Hilbert spaces. In M. Rosenblatt, editor, *Proc. Symposium on Time Series Analysis*, pages 155–169. Wiley, New York, 1963.

[384] E. Parzen. Statistical inference on time series by rkhs methods. In R. Pyke, editor, *Proc. 12th Biennal Seminar, Montreal, Canada: Canadian Mathematical Congress*, pages 1–37, 1970.

[385] W. Paul and J. Baschnagel. *Stochastic Processes: From Physics to Finance.* Springer–Verlag, Berlin, 1999.

[386] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

[387] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

[388] J. Pearl. On the identification of nonparametric structural models. In M. Berkane, editor, *Latent Variable Modelling with Application to Causality*, Lecture Notes in Statistics, pages 29–68. Springer–Verlag, New York, 1997.
Available from `http://singapore.cs.ucla.edu/csl_papers.html`.

[389] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York, 2000.

[390] J. Pearl. Bayesianism and causality, or, why I am only a half-Bayesian. In D. Corfield and J. Williamson, editors, *Foundations of Bayesianism*, Kluwer Applied Logic Series. Kluwer Academic Publishers, Dordrecht, 2001.
Available from `http://singapore.cs.ucla.edu/csl_papers.html`.

[391] H.-O. Peitgen, H. Jürgens, and D. Saupe. *Chaos and Fractals.* Springer–Verlag, New York, 1992.

[392] R. Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51:406–413, 1955.

[393] R. Penrose. On best approximate solutions of linear matrix equations. *Proceedings of the Cambridge Philosophical Society*, 52:17–19, 1956.

[394] V. Periwal. Reparametrization invariant statistical inference and gravity. *Phys. Rev. Lett.*, 78:4671–4674, 1997.

[395] V. Periwal. Geometric statistical inference. *Nucl. Phys. B*, 554:719–730, 1999.

[396] P. Perona and J. Malik. Scale–space and edge detection using anisotropic diffusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[397] M. E. Perskin and D. V. Schroeder. *An Introduction to Quantum Field Theory.* Addison–Wesley, Reading, MA, 1995.

[398] D. A. Pierre. *Optimization Theory with Applications.* Dover, New York, 1986. Original edition published by Wiley, New York, 1969.

[399] T. Poggio and F. Girosi. A theory of networks for approximation and learning. Artificial Intelligence Laboratory Memo 1140, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[400] T. Poggio and F. Girosi. Networks for approximation and learning. In *Proceedings of the IEEE*, volume 78, pages 1481–1497, 1990.

[401] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[402] E. Polak. *Optimization.* Springer–Verlag, New York, 1997.

[403] D. Pollard. *Convergence of Stochastic Processes.* Springer–Verlag, New York, 1984.

[404] D. A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems 1*, pages 305–313. Morgan Kaufmann, San Mateo, CA, 1989.

[405] M. Pontil, S. Mukherjee, and F. Girosi. On the noise model of support vector machine regression. In *Proceedings of Algorithmic Learning Theory 2000 Conference, Sydney, Australia, December 11-13, 1998.*

[406] K. Popper. *Logik der Forschung.* Julius Springer Verlag, Vienna, 1935. English translation: *The Logic of Scientific Discovery.* Hutchinson, London, 1959.

[407] K. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge.* Routledge, London, 1963.

[408] A. Pordt. Random walks in field theory. In H. Meyer-Ortmanns and A. Klümper, editors, *Field Theoretical Tools for Polymer and Particle Physics.* Springer–Verlag, Berlin, 1998.

[409] J. Preskill. Quantum information and computation. A course on quantum computation at Caltech, 1998. Available from `http://www.theory.caltech.edu/people/preskill`.

[410] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, Cambridge, 1992.

[411] F. P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and Other Logical Essays.* Routledge and Kegan Paul, London, 1931.

[412] M. Reed and B. Simon. *Functional Analysis I.* Academic Press, Boston, 1980.

[413] P. Ring and P. Schuck. *The Nuclear Many–Body Problem.* Springer–Verlag, New York, 1980.

[414] B. D. Ripley. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society B*, 39:172–212, 1977.

[415] B. D. Ripley. *Stochastic Simulation.* Wiley, New York, 1987.

[416] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[417] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[418] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

[419] H. Ritter, T. Martinetz, and K. Schulten. *Neuronale Netze.* Addison–Wesley, Bonn, second edition, 1991.

[420] H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proc. Second Berkely Symp. Math. Statist. Probab.*, volume 1, pages 131–148. University of Berkeley Press, Berkeley, 1951.

[421] H. Robbins. An empirical Bayes approach to statistics. In *Proc. Third Berkely Symp. Math. Statist. Probab.*, volume 1, pages 157–164. University of Berkeley Press, Berkeley, 1955.

[422] H. Robbins. The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, 35:1–20, 1964.

[423] H. Robbins. Some thoughts on empirical Bayes estimation. *Ann. Statist.*, 1:713–755, 1964.

[424] C. P. Robert. *The Bayesian Choice.* Springer–Verlag, New York, 1994.

[425] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[426] C. C. Rodriguez. Cross validated non parametric Bayesianism by Markov chain Monte Carlo, 1997. arXiv:physics/9712041.

[427] R. Rojas. *Theorie der neuronalen Netze.* Springer, Berlin, 1993.

[428] K. Rose, E. Gurewitz, and G. C. Fox. Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.*, 65:945–948, 1990.

[429] F. Rosenblatt. *Principles of Neurodynamics.* Spartan, New York, 1962.

[430] H. J. Rothe. *Lattice Gauge Theories.* World Scientific, Singapore, 1992.

[431] T. Roths, D. Maier, C. Friedrich, M. Marth, and J. Honerkamp. Determination of the relaxation time spectrum from dynamic moduli using an edge preserving regularization method. *Rheol. Acta*, 39(2):163–173, 2000.

[432] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

[433] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group. *Parallel Distributed Processing,* volumes 1 & 2. MIT Press, Cambridge, MA, 1986.

[434] P. A. Ruud. *An Introduction to Classical Econometric Theory.* Oxford University Press, Oxford, 2000.

[435] L. H. Ryder. *Quantum Field Theory.* Cambridge University Press, Cambridge, 1996.

[436] L. J. Savage. *Foundations of Statistics.* Wiley, New York, 1954. Also published by Dover, New York, 1972.

[437] H. Scheffé. *The Analysis of Variance.* Wiley, New York, 1959.

[438] M. J. Schervish. *Theory of Statistics.* Springer–Verlag, New York, 1995.

[439] S. J. Schmidt, editor. *Der Diskurs des Radikalen Konstruktivismus.* Suhrkamp, Frankfurt a. M., 1996.

[440] B. Schölkopf, P. L. Bartlett, A. Smola, and R. Williamson. Shrinking the tube: A new support vector regression algorithm. In M.S. Kearns, S. S. Solla, and Cohn D. A., editors, *Advances in Neural Information Processing Systems 11,* pages 330–336. MIT Press, Cambridge, MA, 1999.

[441] B. Schölkopf, C. Burges, and A. Smola, editors. *Advances in Kernel Methods: Support Vector Machines.* MIT Press, Cambridge, MA, 1998.

[442] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2001.

[443] L. L. Schumaker. *Spline Functions: Basic Theory.* Wiley, New York, 1981.

[444] H.-P. Schwefel. *Evolution and Optimum Seeking.* Wiley, New York, 1995.

[445] H. S. Seung. Annealed theories of learning. In *Neural Networks: The Statistical Mechanics Perspective. Proceedings of the CTP–PBSRI Joint Workshop on Theoretical Physics.* World Scientific, Singapore, 1995.

[446] G. Shafer. Constructive probability. *Synthese,* 48:1–60, 1981.

[447] G. Shafer and P. Shenoy. Probability propagation. *Annals of Mathematics and Artificial Intelligence,* 2:327–352, 1990.

[448] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 26:379–423, 623–656, 1948.

[449] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. A framework for structural risk minimization. Technical Report NC–TR_96–032, NeuroColt Technical Report Series, 1996.

[450] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data–dependent hierachies. Technical Report NC–TR_96–053, NeuroColt Technical Report Series, 1996.

[451] P. P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. In G. Shafer and J. Pearl, editors, *Readings in Uncertain Reasoning*, pages 575–609. Morgan Kaufmann, San Mateo, CA, 1990.

[452] N. Shephard and M. K. Pitt. Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667, 1997.

[453] R. Shiller. A distributed lag estimator derived from smoothness priors. *Econometrica*, 41:775–778, 1973.

[454] A. Shleifer. *Inefficient Markets.* Oxford University Press, Oxford, 2000.

[455] B. W. Silverman. Spline smoothing: The equivalent variable kernel method. *Ann. Statist.*, 12:898–916, 1984.

[456] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London, 1986.

[457] D. S. Sivia. *Data Analysis: A Bayesian Tutorial.* Oxford University Press, Oxford, 1996.

[458] J. Skilling. Fundamentals of maxent in data analysis. In B. Buck and V. A. Macaulay, editors, *Maximum Entropy in Action*, pages 19–40. Oxford University Press, Oxford, 1991.

[459] J. Skilling. On parameter estimation and quantified maxent. In W. T. Grandy and L. H. Schick, editors, *Maximum Entropy and Bayesian Methods. Laramie, 1990*, pages 267 –273. Kluwer Academic Publishers, Dordrecht, 1991.

[460] A. F. M. Smith and G. O. Roberts. Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society B*, 55:3–23, 1993.

[461] A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors. *Advances in Large Margin Classifiers.* MIT Press, Cambridge, MA, 2000.

[462] A. J. Smola and B. Schölkopf. From regularization operators to support vector kernels. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*. MIT Press, Cambridge, MA, 1998.

[463] A. J. Smola and B. Schölkopf. A tutorial on support vector machines. Technical Report TR–1998–030, Neuro COLT, 1998. Available from `http//svm.first.gmd.de`.

[464] A. J. Smola, B. Schölkopf, and K-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[465] R. Snieder and A. Tarantola. Imaging of quantum mechanical potentials. *Phys. Rev. A*, 39:3303–3309, 1989.

[466] P. Sollich. Probabilistic interpretation and Bayesian methods for support vector machines. In *Proceedings of ICANN 1999*, pages 91–96, 1999.

[467] P. Sollich. Probabilistic methods for support vector machines. In S. A. Solla, T. K. Leen, and K.-R-Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 349–355. MIT Press, Cambridge, MA, 2000.

[468] P. Sollich. Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning*, 46(1–3):21–52, 2002.

[469] R. G. Staudte and S. J. Sheather. *Robust Estimation and Testing*. Wiley, New York, 1990.

[470] A. Steane. Quantum computing. *Rep. Prog. Phys.*, 61:117–173, 1998. arXiv:quant-ph/9708022.

[471] M. Stone. Cross–validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36:111–147, 1974.

[472] M. Stone. An asymptotic equivalence of choice of model by cross–validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, 39:44, 1977.

[473] M. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13:689–705, 1985.

[474] C. E. M. Strauss, D. H. Wolpert, and D. R. Wolf. Alpha, evidence, and the entropic prior. In A. Mohammad-Djafari, editor, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, Dordrecht, 1992.

[475] K. K. Sung and T. Poggio. Finding human faces with a Gaussian mixture distribution-based face model. In *Recent Progress in Computer Vision*, LNCS Series. Springer–Verlag, New York, 1995.

[476] M. Tan. An inverse problem approach to optical homodyne tomography. *J. Mod. Opt.*, 44:2233, 1997.

[477] M. A. Tanner. *Tools for Statistical Inference.* Springer–Verlag, New York, second edition, 1993.

[478] A. Tarantola. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.* Elsevier, Amsterdam, 1987.

[479] J. R. Taylor. *Scattering Theory.* Krieger, Malabar, FL, 1972.

[480] C. J. Thompson. *Mathematical Statistical Mechanics.* Princeton University Press, Princeton, NJ, 1972.

[481] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22:1701–1762, 1994.

[482] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

[483] A. N. Tikhonov and V. Arsenin. *Solution of Ill-posed Problems.* Wiley, New York, 1977.

[484] J. Timmer. Parameter estimation in nonlinear stochastic differential equations. *Chaos, Solitons, and Fractals*, 11:2571–2578, 2000.

[485] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York, 1985.

[486] G. F. Trecate, C. K. I. Williams, and M. Opper. Finite–dimensional approximation of Gaussian processes. In M.S. Kearns, S. S. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 218–224. MIT Press, Cambridge, MA, 1999.

[487] A. Tversky. Elimination by aspects. *Psychological Review*, 79(4):281–301, 1972.

[488] A. Tversky and D. Kahnemann. The framing of decisions and the psychology of choice. *Science*, 211:453–458, 1981.

[489] J. Uhlig. Bayessches Lernen in der Quantentheorie. Ph.D. Thesis, Universität Münster, 2000.

[490] B. C. van Fraassen. *The Scientific Image.* Clarendon Press, Oxford, 1980.

[491] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry.* North–Holland, Amsterdam, 1992.

[492] V. N. Vapnik. *Estimation of Dependencies Based on Empirical Data.* Springer–Verlag, New York, 1982.

[493] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer–Verlag, New York, 1995.

[494] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[495] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Th. Prob. and Its Application*, 17(2):264–280, 1971.

[496] V. N. Vapnik and A. Y. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):283–305, 1991.

[497] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S–Plus.* Springer–Verlag, New York, 1994.

[498] G. Vico. *De antiquissima Italorum sapientia.* Stamperia de' Classici Latini, Naples, 1858. Originally published in 1710.

[499] W. von der Linden, V. Dose, J. Padayachee, and V. Prozesky. Signal and background separation. *Phys. Rev. E*, 59:6527–6534, 1999.

[500] W. von der Linden, R. Fischer, and V. Dose. Evidence integrals. In R. N. Silver and K. M. Hanson, editors, *Maximum Entropy and Bayesian Methods*, page 443. Kluwer Academic Publishers, Dordrecht, 1996.

[501] W. von der Linden, R. Fischer, and V. Dose. On the importance of α–marginalization in maximum entropy. In R. N. Silver and K. M. Hanson, editors, *Maximum Entropy and Bayesian Methods*, page 229. Kluwer Academic Publishers, Dordrecht, 1996.

[502] E. von Glasersfeld. *Radikaler Konstruktivismus.* Suhrkamp, Frankfurt a. M., second edition, 1998.

[503] J. von Neumann. *Mathematical Foundations of Quantum Mechanics.* Princeton University Press, Princeton, NJ, 1955.

[504] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ, 1947.

[505] G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter. *Numer. Math.*, 24:383–393, 1975.

[506] G. Wahba. Bayesian confidence intervals for the cross–validated smoothing spline. *Journal of the Royal Statistical Society B*, 45, 1983.

[507] G. Wahba. *Spline Models for Observational Data.* SIAM, Philadelphia, 1990.

[508] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Machines*, pages 69–88. MIT Press, Cambridge, MA, 1998.

[509] G. Wahba and S. Wold. A completely automatic French curve. *Commun. Statist.*, 4:1–17, 1975.

[510] P. D. Wasserman. *Advanced Methods in Neural Computing.* Van Nostrand Reinhold, New York, 1993.

[511] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65:499–556, 1993.

[512] P. Watzlawick, editor. *How Real Is Real?* Random House, New York, 1976.

[513] P. Watzlawick, editor. *The Invented Reality.* Norton, New York, 1984.

[514] A. R. Webb. Functional approximation by feed–forward networks: A least–squares approach to generalization. *IEEE Trans. Neural Networks*, 5:363–371, 1994.

[515] S. Weinstein. *The Quantum Theory of Fields*, volume 1. Cambridge University Press, Cambridge, 1995.

[516] S. Weinstein. *The Quantum Theory of Fields*, volume 2. Cambridge University Press, Cambridge, 1996.

[517] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models.* Springer–Verlag, New York, 1997.

[518] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1981.

[519] E. T. Whittaker. On a new method of graduation. *Proc. Edinborough Math. Assoc.*, 78:81–89, 1923.

[520] C. K. I. Williams. Gaussian processes. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks.* MIT Press, Cambridge, MA, second edition, 2002.

[521] C. K. I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20:1342–1351, 1998.

[522] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 514–520. MIT Press, Cambridge, MA, 1996.

[523] P. Wilmott, J. Dewynne, and S. Howison. *Option Pricing: Mathematical Models and Computation.* Oxford Financial Press, Oxford, 1993.

[524] G. Winkler. *Image Analysis, Random Fields, and Dynamic Monte Carlo Methods.* Springer–Verlag, Berlin, 1995.

[525] D. H. Wolpert. A rigorous investigation of 'evidence' and 'Occam factors' in Bayesian reasoning. Working Paper SFI 92-03-013, Santa Fe Institute, 1992.
Available from `http//www.santafe.edu/sfi/publications`.

[526] D. H. Wolpert. On the use of evidence in neural networks. In Stephen José Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 539–546. Morgan Kaufmann, San Mateo, CA, 1993.

[527] D. H. Wolpert, editor. *The Mathematics of Generalization. The Proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning.* Santa Fe Institute, Studies in the Sciences of Complexity. Addison–Wesley, Reading, MA, 1995.

[528] D. H. Wolpert. The existence of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1391–1420, 1996.

[529] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.

[530] D. H. Wolpert and C. E. M. Strauss. What Bayes has to say about the evidence procedure. In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, pages 61–78. Kluwer Academic Publishers, Dordrecht, 1996. An extended version can be found in the Santa Fe Institute Technical report 94-07-001.

[531] S. J. Yakowitz and F. Szidarovsky. A comparison of kriging with nonparametric regression methods. *J. Multivariate Analysis*, 16:21–53, 1985.

[532] A. L. Yuille. Generalized deformable models, statistical physics and matching problems. *Neural Computation*, 2(1):1–24, 1990.

[533] A. L. Yuille and J. J. Kosowski. Statistical physics algorithm that converge. *Neural Computation*, 6(3):341–356, 1994.

[534] A. L. Yuille, P. Stolorz, and J. Utans. Statistical physics, mixtures of distributions, and EM algorithm. *Neural Computation*, 6(2):334–340, 1994.

[535] B. N. Zakhariev and V. M. Chabanov. New situation in quantum mechanics. *Inverse Problems*, 13:R47–R79, 1997.

[536] H. Zhu, C. K. I. Williams, R. Rohwer, and M. Morciniec. Gaussian regression and optimal finite dimensional linear models. Technical Report NCRG/97/011, Aston University, Birmingham, 1997.

[537] S. C. Zhu and D. Mumford. Prior learning and Gibbs reaction–diffusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(11):1236–1250, 1997.

[538] S. C. Zhu, Y. N. Wu, and D. Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8):1627–1660, 1997.

[539] S. C. Zhu and A. L. Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

[540] W. Zhu and H. Rabitz. Potential surfaces from the inversion of time dependent probability density data. *J. Chem. Phys.*, 111:472–480, 1999.

[541] J. Zinn-Justin. *Quantum Field Theory and Critical Phenomena*. Oxford Science Publications, Oxford, 1989.

Index

δ–loss, 62
ϵ–insensitive error, 134, 142, 165

absolute loss, 61
adaptive reference function, 216
adaptive template function, 197–201, *see also* flexible reference function
additive models, 3, 174, 179–181
  generalized, 3, 179–181
annealed field, 342
annealing techniques, 24, 191, 216, 267, 282, 331–332, 349
annealing temperature, 282
apparatus function, 14, 80, 127
approximate fractal, 122–123, 143, 145, 159, 160, 309
approximate invariance, 116–123, 143–145, 276–283
approximate periodicity, 120–122, 126, 143, 155, 156, 276, 279, 280, 285
approximate symmetries, 116–123, 143–145, 276–283
approximately invariant regression, 143–145
auxiliary fields, 221–224, 276, 277, 281, 345
average energy, 267, 273, 279–281, 283

Bayes' theorem, 28–31, 71, 73, 75
Bayesian decision theory, 28, 59–72
Bayesian field theory, 5, 7, 8, 231
Bayesian inverse Hartree–Fock theory (BIHF), 301–306
Bayesian inverse many–body theory, 299–306
Bayesian inverse quantum theory (BIQT), 257–306
Bayesian statistics, 3, 5, 8–50
Bayesian time–dependent quantum theory (BITDQ), 288–293
beliefs, 319
bias–variance decomposition, 125, 138
binomial likelihood, 14, 46, 51, 65, 69
bit number, 21, 25, 326–328, 330
Boltzmann–Gibbs distribution, 25, 43, 331, *see also* generalized canonical ensemble
bootstrap, 67, 185, 189, 209, 292
Brownian motion, 40, 147, *see also* Wiener process

canonical ensemble, 328
  density operator, 259, 260, 279, 300
  for classical systems, 273
  for many–body quantum systems, 299
  for quantum systems, 259, 260, 279, 300
center of mass, 269
classical limit of inverse quantum statistics, 273–275
classification, 3–6, 8, 14, 16, 41, 42, 62, 91, 161–165, 183, 185, 309, 362
cluster regression, 141

clustering, 4, 63, 141, 165
collapse of the quantum mechanical wave function, 258
computer tomography, 127, 257
conditional normalization, 363
conjugate gradient method, 348, 349, 357
connected correlation function, 332, 333, 339, *see also* cumulants
constructivism, 320, 321
controlled prior, 44, 73
cross entropy, *see* Kullback–Leibler entropy
cross–validation, 67, 106, 144, 185, 189, 209–214, 292, 364
  generalized, *see* generalized cross–validation
cumulant expansions, 337
cumulant generating function, 25, 26, 334, 335, 337
cumulants, 25, 26, 330–337
curvature–driven smoothing, 185

data
  dependent part, 4
  finite number of, 264
  independent, 12, 13, 194
  independent part, 4
  large number of, 7
  non–training, 8
  observational, 4, 9, 14, 20, 27, 257, 280, 281, 288, 292, 293, 299, 300, *see also* data, training
  prior, 13, 15
  sampled, 279, *see also* data, training
  sampling, 65
  test, 3, 48
  training, 3, 6, 8, 12, 13, 15, 20, 72, *see also* data, observational
decision trees, 3, 124, 182, 186, 197, 215, 230, 309
decoherence, 258
density operator, 165, 258–260, 279, 300, 310
  for canonical ensemble, *see* canonical ensemble density operator
  for spin systems, 258
design matrix, 175, 178
deterministic annealing, 24
Dirichlet boundary conditions, 352, 356, 358
discontinuities, 119, 120, 221, 223–225, 283, 287
discontinuous template function, 221
discrepancy, 66, 67, 209, 242, 292
dissipation–fluctuation theorems, 336
distribution field, 115–116
downhill simplex method, 191, 349
dummy variables, 173

effective action, 337
effective template function, 220, 233
empirical Bayes approach, 194, 195
empirical risk, 66
empirical risk minimization, 27, 62, 64–72, 96–99, 188, 209
energy, 7, 20–27, 46, 49, 64, 91, 96, 102, 103, 106–108, 110, 115, 130–132, 170, 174, 177–179, 204, 211, 235, 239, 323–325, 328–331, 334, 336, *see also* error functional
  annealed, 341
  annealed prior, 189
  auxiliary prior, 278
  combined, 32
  complete, 111, 128
  component, 232
  conditional, 341
  conditional prior, 189
  hyperparameter, *see* hyperparameter energy

hyperprior, *see* hyperprior energy
joint, 341
joint prior, 189, 190
likelihood, 26, *see also* training energy
mixture, 341
mixture prior, 232
posterior, *see* posterior energy
prior, *see* prior energy and *also* specific prior energy
quadratic, 86
quenched, 341
quenched prior, 189
specific prior, *see* specific prior energy and *also* prior energy
symmetry, 266
training, *see also* energy, likelihood
energy functional, *see* energy and error functional
energy penalty, 270, 272, 275, 283–285
entropy, 22, 24, 25, 33, 34, 43, 328, 330–331
error functional, 27, 64, 66, 67, 70, 71, 85, 86, 89, 92–94, 96, 98, 99, 101, 112, 124, 161, 168, 185, 189, 197, 198, 200, 202, 209, 231, 345–348, 358, *see also* energy
Euclidean action, 328
Euclidean field theory, 5, 20, 26, 216
Euclidean propagator, 351
evidence, 28, 39, 193, 194, 209
expected risk, 59, 66, 96
expert knowledge, 162, 230, 321
expert systems, 3
exponential family, 24, 330, 332
external field, 25, 125, 330, 334, 336, 342
factorizing template function, 182
feedforward neural networks, *see* neural networks
fermions, 299–301, 306, 311
ferromagnet, 242–243
field, 4, 5, 15, 28, 85, 92, 98, 99, 161, 167, 188, 191, 211, 217, 334–337, 339, 342, 349
annealed, *see* annealed field
auxiliary, *see* auxiliary field
distribution, *see* distribution field
external, *see* external field
general, *see* general fields
hyper–, *see* hyperfield
likelihood, *see* likelihood field
log–likelihood, *see* log–likelihood field
magnetic, *see* magnetic field
quenched, *see* quenched field
scalar, *see* scalar field
square–root, *see* square–root field
field strength tensor, 204
field theories, 5, 6, 203, 204
filter operator, 83, 217, 221, 223
filtered difference, 217–219, 222–224, 227, 248, 276, 277, 281, 282
Fisher information, 114, 115, 317
flexible reference function, 126, 197–201, *see also* adaptive template function
free energy, 7, 23–26, 32, 326, 329, 330
likelihood, 26
posterior, 26
prior, 42
free field, 309, *see also* Brownian motion and Wiener process
generalized, 309, *see also* Gaussian process
free scalar field, 351
Frequentist statistics, 28, 40, 64–69, 71, 96, 185, 188, 209,

210, 314, 320
functional derivative, 89, 112, 134, 188, 196, 198, 224, 262, 263, 267, 268, 275, 277, 278, 290, 301–303, 331, *see also* variational derivative
functional integral, 27, 41, 265, *see also* path integral
  Gaussian, 27
  non–Gaussian, 27
fuzzy methods, 230, 232, 319

gauge fields, 204
Gauss–Seidel iteration, 348, 358
Gaussian mixture prior, 8, 231–257, 272, *see also* mixture of Gaussian processes
Gaussian mixture regression, 141
Gaussian prior factor, *see* Gaussian process
Gaussian process, 5–8, 18, 19, 42, 66, 73, 85–164, 190, 198, 220, 229, 265–267, 269, 272, 275, 276, 300, 309, 310, 316, 350
Gaussian process prior, *see* Gaussian process
Gaussian regression, 126–137
Gaussian relaxation, 354
general fields, 112–114
generalized canonical ensemble, 25, 43, 331, *see also* Boltzmann–Gibbs distribution
generalized cross–validation, 214
generalized free field, 6, 309, *see also* Gaussian process
generalized linear models, 172
generalized method of moments (GMM), 17
generating function, 332–339
  for cumulants, *see* cumulant generating function
  for moments, *see* moment generating function
genetic algorithms, 191, 216, 349
Gibbs distribution, *see* Boltzmann–Gibbs distribution
Gibbs ensemble, *see* Boltzmann–Gibbs distribution
Gibbs random field, 20
gradient, 36, 45, 87, 88, 90, 91, 93, 95, 99, 111, 119, 182, 191, 261, 268, 336, 348, 350, 358
gradient algorithm, 36, 45, 87, 91, 93, 95, 101, 102, 110, 225, 348, 349, 351, 354, 355, 358
grand canonical ensemble, 328
grapical models, 3, 141, 324
Grassmann variables, 343
Green's function, 351, 352
ground state energy, 115, 264, 270, 289

Hamiltonian, 115, 197, 257, 259–261, 283, 288, 300, 304, 310
  Hartree–Fock, 301
  many–body, 299–301
  nonlocal, 288
  real, 289, 303
  two–body, 300
hard constraints, 8, 192, *see also* non–negativity and normalization
Hartree approximation, 182, 197
Hartree–Fock approximation, 72, 182, 197, 301, 302, 306, 312
Hartree–Fock likelihood, 302
Hartree–Fock–Bogoliubov approximation, 197
Hessian, 36, 37, 39, 91–92, 95–96, 101, 102, 106, 110, 111, 114, 176, 191, 192, 262, 268, 336, 350, 351, 358, 360, 361

hidden states, *see* variables, hidden
high temperature limit, 24, 240, 242, 264, 332
high temperature solution, 242, 247, 256
hinge functions, 180
hints, 185, 186
homotopy methods, 191, 216, 331–332, 349
Huber's loss function, 142
hyperfields, 199, 202–204, 216–223, 276, 277, 282, 283, 287, 310, 311, 320
hyperparameter energy, 198, 207, *see also* hyperprior energy
hyperparameters, 7, 8, 18, 19, 67, 112, 121, 124, 144, 167, 187–227, 229, 250, 251, 254, 265, 280, 289, 292, 310, 317, 319
  discrete, 229
hyperprior energy, 204, 239, *see also* hyperparameter energy
hypothesis testing, 64, 65, 68, 69

image completion, 187, 249–257, 310
infinitesimal translations, 116–120, 143, 147, 266
influence matrix, 132, 134, 176
input noise, 17, 19, 75–78, 80, 81, 84, 318
instrument function, 14, 80, 127
interacting fields, 6
interaction terms, 173, 174, 180, 186, 203
invariant group measure, 82, 83
inverse Hartree–Fock equation, 303
inverse problems, 3
  classical, 14, 42, 44, 80, 127, 211, 257
  in quantum theory, 4, 257–306, *see also* inverse quantum theory
inverse quantum theory, 8, 257–306
inverse scattering theory, 257, 264
inverse spectral theory, 257, 263, 264
inverse temperature, 24–27, 184, 188, 207, 232, 239, 245, 246, 336
inverse Wishart distribution, 208

Jacobi iteration, 348, 358
Jacobi matrix, *see* Jacobian
Jacobian, 90, 112, 168, 196
Jensen's inequality, 63

Kerridge inaccuracy, 22, 40
kinetic energy, 261, 288, 300, 351
Kuhn–Tucker conditions, 88
Kullback–Leibler distance, *see* Kullback–Leibler entropy
Kullback–Leibler entropy, 23, 25, 43, 60, 63, 110

Lagrange multiplier, 24, 25, 43, 45, 46, 85–95, 99, 113, 131, 161, 162, 168, 177, 222, 232, 267, 330–332
Lagrange multiplier function, *see* Lagrange multiplier
Lagrange parameter, *see* Lagrange multiplier
Landau–Ginzburg models, 230
Laplace's approximation, *see* Maximum A Posteriori Approximation
lattice field theories, 6
learning matrix, 346, 348–357
Legendre transform, 334, 337
Lie group, 82, 118, 266
likelihood
  normalization, 29, 89–91, 94–95, 112, 239, 246
likelihood field, 93
likelihood model of quantum theory, 85, 258–263, 310

linear regression, 4, 174–179
linear trial spaces, 171–174
linguistic variables, 230, 232
local Gaussian mixture prior, 248–249
local mass, 203–204
local reference function, 216–221, 248–249
local template function, 216–221, 248–249
log–likelihood, 4, 21, 45, 66, 85–89, 92, 97, 106, 107, 114, 232, 262, 292
  non–quadratic, 134
  shifted, 162
log–likelihood field, 92
log–loss, 59–61, 63–64, 70, 71, 96, 107, 141, 245
logistic regression
  binary, 173
  cumulative, 174
  multinomial, 174
  ordinal, 174
logit, 173
loss, 15, 59–72, 142
  ϵ–insensitive, *see* ϵ–insensitive error
  absolute, 61
  delta, *see* δ–loss
  general quadratic, 61
  Huber's function, *see* Huber's loss function
  logarithmic, *see* log–loss
  quadratic, *see* quadratic error or squared–error loss
  quadratic ϵ–insensitive, *see* quadratic ϵ–insensitive error
  squared–error, *see* squared–error loss
  zero–one, 62, 161
low temperature limit, 24, 40, 232, 240–242, 332
low temperature solution, 242, 247, 251, 255
magnetic field, 125, 330, 334
many–body Schrödinger equation, 301
MAP, *see* Maximum A Posteriori Approximation
Markov random field, *see* Gibbs random field
mass, 99, 106, 108, 110, 235, 238, 261, 273, 293–295, 305, 350–352, 354, 359, 361
massive relaxation, 103, 106, 110, 111, 233, 350–353, 358
Maximum A Posteriori Approximation, 27, 36–40, 48, 50, 63–66, 68, 70, 71, 90, 91, 131, 132, 161, 162, 167, 188, 191–197, 222, 267–269, 312
maximum entropy, 24, 25, 43, 331
maximum entropy prior, 43, 44
maximum likelihood approach, 64, 65, 71, 261
Maximum Likelihood Approximation, *see* maximum likelihood approach
mean field approximation, 258
mean field equation, 243
mean field theory, 72, 162
measured prior, 44, 73–75
median, 61, 62
method of steepest descents, *see* Maximum A Posteriori Approximation
mixing of templates, 215
mixing template functions, 230
mixture of filtered differences, 218, 277
mixture of Gaussian processes, 19, 229–257, 266–268, 271, *see also* Gaussian mixture prior
mode, 48, 49, 62
modified Bessel functions, 351, 353
moment expansions, 337
moment generating function, 25, 26, 333

moments, 25, 26, 330, 332, 333, 335–337
monotonicity, 41, 99, 113, 116
Monte Carlo methods, 36, 71, 161, 188, 192, 209, 216, 265, 292, 324, 337
Moore–Penrose inverse, 135, 136, 143, 169, 262, 263, 290, 355
multimodal energy surface, 229
Multiple Reduction Copy Machines (MRCM), 122
mutual information, 33–36, 38, 60, 315

negative binomial likelihood, 65, 70
Neumann boundary conditions, 352
neural networks, 3, 4, 180, 183–186, 215, 230
  multilayer, 183–186
  one-layer, *see* perceptron
Newton method, 91, 268, 348–350, 358
noise induced prior, 75–84, 93
non–negativity, 6, 21, 27, 41–46, 78, 93, 99, 112–114, 116, 161, 198, 349, 358
  and monotonicity, 41, 99, 116
non–quadratic potentials, 8, 224–227
nonparametric boosting, 198
normal equation, 175
normalization, 6, 23, 27, 29, 33, 34, 40–46, 71, 74, 78, 83, 86–91, 93–97, 99, 102, 106, 107, 112–116, 131, 139, 141, 161, 162, 168–170, 185–192, 197, 198, 202–207, 218, 220, 222–224, 232, 239, 246, 262, 264, 283, 310, 323–343, 345, 347, 349, 358, 362
  in L_2-norm, 114
  of eigenfunctions, 262, 264
  of likelihoods, 29, 89–91, 94–95, 112, 239, 246
  of predictive density, 139
  of priors, 33, 187–191, 202–207, 220, 222–224, 246, 283, 310
normalization constraint, *see* normalization
normalization energy, 207
null hypothesis, 65, 69

observation equations, 29
Occam's razor, 67–70
optical flow field, 187, 250, 310
outliers, 143, 224, 225
output averaging, 75, 78, 79
output noise, 17, 19, 73–75, 77, 78, 81, 82, 84, 116, 142, 318
overfitting, 67, 112, 144, 185, 192, 233, 237, 271, 280, 292

P–value, 65, 70
pairing fields, 197
particle mass, 261, 273, 293–295, 305
partition sum, 23, 26, 324–325, 342
Parzen's kernel, 99, 352
path integral, 39, *see also* functional integral
penalized maximum likelihood approach, 66, 71
perceptron, 183, 184
piecewise smooth, 224
plug–in estimate, 64, 66, 210
pointer states, 258
Poisson likelihood, 14, 127, 165
posterior, 9–84
posterior control, 313–321
posterior energy, 26, 27, 130, 131, 168
posterior mode, 48, 49, *see also* Maximum A Posteriori Approximation
posterior state of knowledge, 13, 20
prediction error, 210

predictive density, 9–84
prior, 9–84
  controlled, *see* controlled prior
  Gaussian, 8, 85–164, *see also* Gaussian process
  Gaussian mixture, *see* Gaussian mixture prior
  improper, 208, 317
  local Gaussian mixture, *see* local Gaussian mixture prior
  maximum entropy, *see* maximum entropy prior
  measured, *see* measured prior
  noise induced, *see* noise induced prior
  non–quadratic potentials, 8, 224–227
  noninformative, 208, 317
  normalization, 33, 187–191, 202–207, 220, 222–224, 246, 283, 310
  Radial Basis Function (RBF), *see* Radial Basis Function (RBF) prior
  smoothness, *see* smoothness prior
  structural, *see* structural prior
prior energy, 26, 40, 43, 100, 116, 117, 121, 123, 124, 126, 128, 130, 170, 178, 181, 189, 203, 204, 208, 217–220, 222, 223, 227, 230, 267, 277, 278, 281–283, 300
prior mode, 49
prior relaxation, 103, 106, 233, 348, 351, 358
product ansatz, 181–182
projection matrix, 176
projection pursuit, 3, 63, 180, 182–183, 186, 269, 309
prospect theory, 59
Pseudo Maximum Likelihood Approximation, 71

quadratic ϵ–insensitive error, 142
quadratic error, 66, 86, 96, 266
quantum Anti–Zeno effect, 291
quantum field theory, 5, 39
quantum Zeno effect, 291, 296
Quasi Maximum Likelihood Approximation, 71
quasi–Newton methods, 268, 348, 349, 357
quenched field, 342

Radial Basis Function (RBF) prior, 43, 143, 148, 151, 152, 265, 271, 291, 354
Radial Basis Functions (RBF), 179, 182, 354
rational decisions, 319
reference function, 43, 68, 126, 215, 218–220, 222, 273, 280, 281, 316, 317, *see also* template function
regression, 60, 126–145
  cluster, *see* cluster regression
  Gaussian mixture, *see* Gaussian mixture regression
  linear, *see* linear regression
  logistic, *see* logistic regression
  projection pursuit, *see* projection pursuit
regression function, 4, 14, 16–18, 37, 43, 45, 60, 76, 79–81, 102, 106–108, 110, 118, 122, 123, 126, 127, 130, 141–148, 151–153, 156, 158, 160, 185, 199–201, 216, 233, 235, 239, 249, 269, 316, 318
regularized error functional, 27, 40, 64, 70, 71, 85, 86, 89, 92–94, 209, 231
relative entropy, *see* Kullback–Leibler entropy
replica approach, 183, 343
replica method, *see* replica approach
resolvent, 351
ridge regression, 180

Ritz method, 172
robust statistics, 142, 143, 225
rotation, 118, 187, 197, 205, 215, 250, 254, 310

saddle point approximation, *see* Maximum A Posteriori Approximation
scalar field, 99, 323
scaling, 122, 187, 197, 208, 215, 254, 310
Schrödinger equation, 115
self energy, 338, 339
simulated annealing, 24, 191, 216, 282, 283, 286
Slater determinant, 300, 302, 304
smoothing splines, 3
smoothness prior, 19, 93, 100–105, 108, 118, 120, 141, 143, 150–152, 157, 187, 203, 221, 235, 265–267, 276, 279–281, 283, 319
Sobolev distance, 265
specific prior energy, 42, 45, 112, 117, 202
spin glasses, 183, 190, 325
spin systems, 258
square–root field, 114–115
squared–error loss, 60, 61, 141, 185, 245
statistical field theory, 5, 26, 32
structural prior, 12, 15, 18, 72, 74, 75, 319
Student-t distribution, 208
subjective probability elicitation, 319
support vector machine, 3, 134, 142, 165, 209
support vector regression, 134, 142, 165
switching between filtered differences, 218, 222, 281, 282
system equations, 29

template function, 122, 124–126, 130, 168, 170, 172, 195, 197–201, 203, 208, 215–217, 219, 221, 223, 230, 232–235, 241, 242, 247, 251, 253, 265, 281, *see also* reference function
training energy, 126, 127, 167, 175
truncated Radial Basis Function prior, *see* Radial Basis Function (RBF) prior

unnormalized likelihood, 161
unsupervised learning, 4

Vapnik–Chervonenkis–dimension, 315
variable metric methods, 268, 348
variables
  continuous, 5, 7, *see also* variables, real
  dependent, 9–20
  discrete y, 4
  hidden, 9–20, 28, 29, 46, 47, 64, 74, 75, 85, 112, 167, 188, 191, 194, 239, 313–315, 318–320
  independent, 9–20
  infinite number of, 4
  latent, *see* variables, hidden
  multidimensional, 4
  real, 4, 8, *see also* variables, continuous
variational derivative, 87, 116, *see also* functionnal derivative
variational method, 6, 115, 167, 168, 309
vector field, 118
virtual training examples, 186
von Neumann series, 83

Wiener process, 144, 147, *see also* Brownian motion

zero mode, 74, 76, 84, 91, 92, 96, 120, 350, 351, 354